July 2010

The Perception of Natural, Cell Phone, and Computer-Synthesized Speech During The Performance Of Simultaneous Visual-Motor Tasks

Nirmal Kumar Srinivasan

University of Nebraska at Lincoln, nirmal.srinivasan@huskers.unl.edu

THE PERCEPTION OF NATURAL, CELL PHONE, AND COMPUTER-SYNTHESIZED SPEECH DURING THE PERFORMANCE OF SIMULTANEOUS VISUAL-MOTOR TASKS

by

Nirmal Kumar Srinivasan

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the of Degree Doctor of Philosophy

Major: Intradepartmental Area of Human Sciences

(Communication Disorders)

Under the Supervision of Professor Thomas. D. Carrell

Lincoln, Nebraska

August, 2010

THE PERCEPTION OF NATURAL, CELL PHONE, AND COMPUTER-SYNTHESIZED SPEECH DURING THE PERFORMANCE OF SIMULTANEOUS VISUAL-MOTOR TASKS

Nirmal Kumar Srinivasan, Ph.D.

University of Nebraska, 2010

Advisor: Thomas. D. Carrell

This study investigated the influence of top-down and bottom-up information on speech perception in complex listening environments. Specifically, the effects of listening to different types of processed speech were examined on intelligibility and on simultaneous visual-motor performance. The goal was to extend the generalizability of results in speech perception to environments outside of the laboratory. The effect of bottom-up information was evaluated with natural, cell phone and synthetic speech. The effect of simultaneous tasks was evaluated with concurrent visual-motor and memory tasks. Earlier works on the perception of speech during simultaneous visual-motor tasks have shown inconsistent results (Choi, 2004; Strayer & Johnston, 2001). In the present experiments, two dual-task paradigms were constructed in order to mimic non-laboratory listening environments. In the first two experiments, an auditory word repetition task was the primary task and a visual-motor task was the secondary task. Participants were presented with different kinds of speech in a background of multi-speaker babble and were asked to

repeat the last word of every sentence while doing the simultaneous tracking task. Word accuracy and visual-motor task performance were measured. Taken together, the results of Experiments 1 and 2 showed that the intelligibility of natural speech was better than synthetic speech and that synthetic speech was better perceived than cell phone speech. The visual-motor methodology was found to demonstrate independent and supplemental information and provided a better understanding of the entire speech perception process. Experiment 3 was conducted to determine whether the automaticity of the tasks (Schneider & Shiffrin, 1977) helped to explain the results of the first two experiments. It was found that cell phone speech allowed better simultaneous pursuit rotor performance only at low intelligibility levels when participants ignored the listening task. Also, simultaneous task performance improved dramatically for natural speech when intelligibility was good. Overall, it could be concluded that knowledge of intelligibility alone is insufficient to characterize processing of different speech sources. Additional measures such as attentional demands and performance of simultaneous tasks were also important in characterizing the perception of different kinds of speech in complex listening environments.

ACKNOWLEDGEMENTS

My doctoral training has been an extremely enriching learning experience and I would like to take this opportunity to thank the numerous people who have contributed and helped me throughout this journey. My greatest gratitude to them may not be enough to reflect their contributions to this dissertation and my academic and professional growth.

First and foremost, I would like to acknowledge my mentor, Dr. Thomas Carrell, for providing an opportunity to work under his guidance. He has been a constant source of support throughout my MS and PhD study. He is a great mentor and a very active person who cares greatly about the success of his students. His valuable support, guidance, supervision, and motivation during my PhD program influenced and shaped this dissertation work. His scientific expertise and acumen has and will always be a constant source of inspiration for me. He has provided me with complete independence to learn from my mistakes which has helped me to grow as a better researcher. His honor, patience, work ethics, and integrity will serve as the base line performance I will strive for as I continue to grow in my research career.

My greatly indebted to Dr. Jordan Green, Dr. Newell Decker, and Dr. Cal Garbin for consenting to serve as my dissertation committee members. I appreciate their feedback, suggestions, and guidance during this research through their technical knowledge and expertise. I extend my special thanks

to Dr. Jordan Green and Dr. Newell Decker, members of my dissertation reading committee, in carefully evaluating this dissertation and their feedback for improving this dissertation. I am indebted to Dr. Jordan Green for introducing me to the world of Speech Science; his dedication to the field ignited me to investigate beyond the surroundings and look at the big picture.

My heartfelt acknowledgements to Barkley Memorial Center and especially to Dr. John Bernthal for their continuous support throughout my graduate school. My special thanks to the past and present doctoral students in Barkley Memorial Center: Iggy, Antje, Sangsook, Nori, James, Wendy, Dave, Kelly, Sarah, Jill, Katy, Anusha, Shari, Chris, Kelly, and Trisha. A special thanks to my long time room-mate and a good friend, Vettrivel, for his inputs during experimental design, data collection, and data analysis. I would also like to thank my numerous other friends at Lincoln and other parts of the world for their continuous support and encouragement during this dissertation.

I have gained benefit from many other teachers between then and now, and I feel (in most cases belatedly) grateful to them all. At the risk of unfair omissions resulting from poor memory, I would like to mark my gratitude to Mary Evelene, Kokila, Sharadha, Jikki, Balasubramanian, Xavier, Natarajan, Rex Johnson, Parvat Kumar Raut, Senthil Kumar, and Harshvardhan Gupta.

This accomplishment would never have been possible without the inspiration of my parents, Srinivasan and Mythili. They taught me work ethics, patience, perseverance, and principles and never had any doubts in me during the course of this entire dissertation. You are the best in this world. I would also like to thank my other family members: Sairam, Kamesh, Prashanthi, Arthi, Uma, Arun, Adithya, and Chirag for their love and support.

Last, but not the least, I would like to thank Dharini for her continuous support and love throughout the dissertation process. Without her help, finishing up this dissertation would have been close to impossible.

TABLE OF CONTENTS

FIGURE CAPTIONS

TABLES

CHAPTER I

Introduction

The present work focuses on fundamental speech processing capabilities that apply to difficult speech signals in complex listening and working environments. While these listening tasks sound difficult and unusual, they are probably more numerous than occasions of listening to speech in conditions of laboratory-quiet. Examples of **difficult speech** signals include: Synthetic speech (as found in GPSs), cell phone speech, speech in noise, Voice Over Internet Protocol (VOIP) speech, speech from speakers with a different dialect than that of the listener, and speech from speakers who have voice, articulation, or fluency problems. Examples of **complex listening situations** include: Speech in noisy spaces, speech in classrooms and auditoriums with significant reverberation, and speech in unusual spaces such as underwater where the nitrogen in the air is replaced with helium. Examples of listening to speech while performing **simultaneous tasks** include: speech on the factory floor, speech while driving or flying, and speech during participation in sports. Several studies over the past few decades have demonstrated effects of these conditions on the way we hear and speak as well as the way we perform on simultaneous tasks. Some of these tasks will be described in greater detail later in this chapter.

Speech Perception. Speech perception has been characterized as a complex process with many intermediate stages and the possibility of interaction between the stages (Elman and McClelland, 1984). Understanding speech is more complex than matching a sequence of sounds with an internal template of symbols or meanings. Speech does not have a one-to-one correspondence with linguistic units such as phones, phonemes, and morphemes. Stated more generally, there is no set of sounds that always indicate that a particular phoneme is present nor is there a set of sounds that always indicate that a particular phoneme is not present. This is known as the problem of acoustic-phonetic invariance. In addition, for most phonemes it is impossible to mark a time at which one phoneme ends and the next one begins. This is known as the problem of segmentation. These problems make modeling speech perception difficult and many creative theories have been developed to address these difficulties. Their goals have been to describe how sound is reliably converted into a sequence of linguistic symbols (acoustic-phonetics), or to convert articulation into linguistic symbols (articulatory-phonetics). The speaker's and listener's goals are to transfer thoughts and meanings in one brain to another using the physical world to connect the speaker and listener via an acoustical channel. Measures of intelligibility are used to reflect the quality of this connection.

However, there are other aspects of speech that are communicated between a speaker and a listener (e.g., talker identity, dialect, sincerity, and

gender) and they require different measures. *Non-linguistic* aspects of speech such as these have been studied, but to a lesser extent than linguistic-based intelligibility. More importantly, the way that linguistic and personal aspects of speech **work together** to improve intelligibility have also been shown to be important (Bradlow, Nygaard, & Pisoni, 1999; Pisoni, 1994) but this interaction of linguistic and personal aspects of speech have only rarely been studied.

Another aspect of the perception of speech has received probably the least attention of all. This is the interaction between speech perception and the performance of concurrent tasks (such as conversing while driving, cooking, or operating machinery). It has been shown that the execution of a simultaneous visual or motor task can influence the perception of speech and, conversely, that the quality of speech can influence performance on visual or motor tasks (Choi, 2004; Gordon, 1993; Luce, Feustel, & Pisoni, 1983).

The interactions between speech perception and non-linguistic aspects of speech demonstrated that a complete theory of speech perception must take into account non-linguistic sounds that are contained in the speech signal. Similarly, the interaction between speech perception and simultaneous visual-motor tasks demonstrated that a complete theory of speech perception must take into account the entire environment that the listener is in – including any unrelated visual and motor tasks. Taking these external influences into account is likely to change the current models of

speech perception. In the present work we examine the effects of signal quality and simultaneous visual-motor tasks on speech perception. The results of this work should then make it possible to construct models of speech perception that are more complete and will have broader application. One good example of the practical usefulness of this knowledge and broader applicability might be understanding the interaction between speech communication and driving.

Although this work used no driving or driving simulation in the experimental methodology, the applicability of the present work to driving with poor-quality cell phone communications is clear.

Driving is a task that is frequently conducted while speaking and listening. Along with the theoretical motivations just mentioned, human-machine interaction while driving was an important motivation for investigating the relationship between communication and simultaneous visual-motor tasks. For example, driving is a highly complex task that requires most sensory modalities, creates a significant cognitive load, and requires rapid reaction times. It is a high-stakes task whose improper performance can lead to injury or even death. Although the act of driving itself was not studied in the present work, another visual-motor task, the *pursuit rotor task*, was used to simplify interpretation of the results. The pursuit rotor is based on the work of Snoddy (1926). This task has been used extensively since that time with considerable variation of stimuli, apparatus

and subject. If significant results are found with the pursuit rotor task in the present work, then the present methodologies should be extended and investigated in real-world driving and driving simulators.

Listening to speech while concurrently doing another task is more common than listening to speech alone (Tun & Wingfield, 1994). A modern example of this situation is listening to cell phone speech and driving. Over the last 10 years, cell phones had gained a substantial world wide acceptance. In US alone, there were an estimated number of 270 million cell phone subscribers as of December, 2008. The total number was expected to surpass 300 million subscribers by the end of 2009 (CTIA, 2008). This increase in the number of cell phone users has increased the number of people who simultaneously talk and drive. For example, surveys indicate that 85% of cell phone owners use their phone occasionally while driving (Goodman, Bents, et al., 1999). Moreover, 27% of the people use their phones half of their total trip time or more (Goodman, Tijerina, Bents & Wierwille, 1999).

Measurement of speech quality. The major goal of any communication device or vocoding algorithm is to maximize efficiency in everyday communication. However, the ability to measure the efficiency of communication can be difficult to accomplish because of the multidimensional nature of visual, motor, and auditory tasks. Traditionally, intelligibility measures have been used to measure the effectiveness communication. However, listener effort, naturalness, talker identity,

attention, and memory are also important characteristics of the communication situation. Despite these additional characteristics of speech, the communication function had been frequently been based entirely on intelligibility. Intelligibility has been measured using either behavioral testing methods, which are typically based on the recognition of speech units such as phonemes or syllables or has been mathematically estimated by measuring the transmitted speech energy at different frequency bands

Measurement. The early work on telephonic intelligibility measures largely influenced research approaches in evaluation of quality of signals processed through the communication channels. However, the results predicted from both behavioral measures and mathematical estimations have not performed well at predicting real life performance. Although a variety of speech recognition tests have been developed over the years and are available for research use, criticism of these recognition tests have been growing over the years. This is due to the lack of sensitivity and reliability of these recognition tests for evaluating the efficiency of the communication channels. One of the reasons for the speech recognition tests to be criticized might be related to their simplicity of test procedures as opposed to the complexity of the listener's typical listening environment. Most of these speech recognition tests and speech perception theories were developed in a quiet noise-free laboratory environment under strictly controlled experimental conditions. Moreover, the stimuli used in the development of these theories were free

from reverberation, noise, and distortion. However, scenarios faced in everyday world are acoustically complex. In everyday situations, speech is generally mixed with noise and noise one listens to speech while doing other simultaneous tasks. Understanding speech is more challenging than recognizing a sequence of sounds, and it requires different analyses and processes.

Steps postulated in speech perception. Decoding an acoustic signal presented at the cochlea starts by transducing the auditory signal into firing patterns of auditory neurons in the form of outer and inner hair cell deflections. This neutrally encoded message at the sensory level invokes a series of processes at phonetic, semantic, syntactic, and lexical levels. Each of these invoked processes enhances the presented acoustic signal at different levels. Therefore, even if an entire phoneme is missing or replaced by a non-speech sound, listeners would be able to understand the signal present without noticing the absence or change of the phoneme (Warren, 1970). Hence the changes in the presented signal due to the addition of noise, masking, or other forms of degradation such as delay or jitter may not be well documented by measuring the performance in intelligibility tests. Other higher level cognitive processes such as phonological knowledge, syntactical structure, and semantic predictability helps in extracting the available acoustic cues for the perception of presented speech.

Traditional methods used to measure intelligibility of presented speech seem to be insufficient in characterizing the performance of speech perception models. Researchers have understood these limitations and have added additional dependent measures such as reaction time, dual-task performance, subjective evaluation, brain imaging, and subjective evaluations to better understand the process of speech perception. Listener's subjective evaluation on the quality of signal presented is too difficult to interpret and compare between different listeners due to inter-subject variability and reliability. Inability to use real speech sounds while measuring evoked potential restricts the researchers in having evoked potential as a major objective physiological measure in evaluating perception of speech. Technologies have been developed to a greater extent to use brain imaging as a tool to study speech perception. However, the higher cost and availability of imaging techniques compared to other measurements restricts the use of brain imaging in a broader spectrum.

Researchers have made extensive use of reaction time measurement and dual-task performance techniques to understand mental processes and human performance. Reaction time has been used as a behavioral measure for processing speed and capacity demands (Donders, 1969; Posner, 1978). Reaction time has also been extensively used as an index of mental processing demands imposed in a given speech test (Gatehouse & Gordon, 1990; Pisoni & Tash, 1974; Pratt, 1981). Reaction time measures the time

taken by the listener to respond to the presented auditory stimulus. It provides a measure of assessing the differences in the time taken by different participants for the perception of the stimulus presented. It is generally assumed that the participant takes longer time to respond when additional perceptual processing for the presented stimulus is required. Therefore, differences in reaction time can be used as an indicator for measuring the difficulty in listening to speech with higher reaction time pointing towards higher difficulty in perceiving the speech stimulus presented.

In many types of perceptual-motor tasks, there is a tradeoff between how fast a task can be performed and how many mistakes are made while performing the task. This is known as the speed-accuracy tradeoff. When this occurs, a listener can either perform the task rapidly with many errors or slowly with few errors. When asked to perform a task as well as possible, listeners normally apply various strategies that may optimize speed, optimize accuracy, or jointly optimize both speed and accuracy. Hence the two conditions fast with less accuracy and slow with more accuracy, cannot be compared either on the basis of either speed or accuracy because faster speed on the task by a listener could have been at the cost of more errors (Sawusch, 1996). The participants behave as if they were given different sets of instructions. For this reason, the relationship between speed and accuracy needs to be mapped out in situations where a speed-accuracy tradeoff exists. Also, the differences between two different groups of listeners cannot be

interpreted when speed-accuracy tradeoff exists (Pike, McFarland, & Dalgleish, 1974; Posner, 1978). Moreover, a serial task and a parallel task can have the same reaction time (Townsend, 1990). Hence reaction time requires careful consideration during experimental design to avoid these problems.

Another approach that has been employed for measuring mental task load is to use dual-task paradigms. This methodology allows measuring the limitations on perceptual and cognitive capacity while performing multiple simultaneous tasks. Dual-task methods have been extensively used by NASA (National Aeronautics and Space Administration) and the FAA (Federal Aviation Administration) to test pilot's performance under different task demand levels and performance during simultaneous tasks. In speech perception studies, dual-tasks have been used to measure the increase in mental processing demands for perceiving speech due to the introduction of a secondary distracting task. Also, the interactions between speech perception task and the secondary distractor task could be measured. In a typical dual-task paradigm, listening is the primary task and a simultaneous secondary task is added to increase the overall processing demands of the tasks together. While doing two tasks simultaneously, reduction in the performance on the secondary task is considered to be an indicator of increase in the processing demand of primary listening task to the limitation of processing capacity (Kahneman, 1973).

Other resource theories consider humans as a multiple channel processor each with its own individual capacities and sharing mechanisms for these multiple processes (Novan & Gopher, 1979). Wickens' multiple resource theory proposed that the humans did not have one single information processing source that could be tapped, but several different pools of resources that could be tapped simultaneously. Depending on the nature of the task, these resources may have to process information sequentially if the different tasks require the same pool of resources, or can be processed in parallel if the task requires different resources. Overall, it was consistently found that intelligibility measures did not reflect the change in signal quality, rather increased processing demands and listening efforts characterized the secondary task performance.

Another way of modeling attention is based on the way tasks are learned rather than how much overlap there is between the resources demanded by each task. The most popular of these is the **Automatic versus Controlled processing theory**. Cognitive processes are broadly classified into automatic processes and controlled processes (Schneider & Shiffrin, 1977). Automatic processes tend to be over-learned and allow the individual to rapidly perceive and identify external stimuli. Controlled processes are used when individuals have little training searching for a stimulus. For example, one immediately recognizes a stop sign, even in a cluster of many other signs. Whereas, a sign denoting a recreation area in an unfamiliar state

is likely to require significant time to identify in a similar cluster. In this example, the stop sign is found using automatic processing and the recreation sign is found using controlled processing.

Automatic processing and controlled processing are two qualitatively different forms of processing that provide different benefits based on the nature of the tasks. It is highly difficult for a single process to provide both the fast learning of controlled processing and the high speed parallel processing of automatic processing (Schneider & Chein, 2003). Not only is it likely that these two modes of attention are more efficient than either one alone would be, some researchers have gone as far as to suggest that there are important survival advantages for this capability.

With controlled processing, the fundamentals of new skills can be acquired quickly; however high effort is required and it can operate only on small number of stimuli at a time. In addition, the slow, resource limited nature of controlled processing will be a disadvantage if the task requires the coordination of many sensory and motor inputs simultaneously.

Despite the time and effort required to acquire automatic processing, it has the advantages of: 1) allowing many processes to occur in parallel, 2) being robust under stress, and 3) allowing rapid action.

It is important to consider additional measures such as listening effort when evaluating the effectiveness of any variation from a natural speech signal. As listeners, we now perceive wider varieties of speech that in the

past, due to the current technology which increasingly relies on voice communication. Perceptual performance is a result of complex mental processes that reflects not only the quality of the signal transmitted but also other mental process required to achieve a given performance level. These mental process paradigms are valuable in terms of modeling the perceptual processes. Hence, methods that can reflect this entire effort are an important tool to evaluate the effectiveness, and the quality of the speech presented and the entire environment that the individual tasks are performed concurrently.

A major goal of the present research is to determine the effect of concurrent simultaneous tasks on perception and performance with different kinds of speech. It is proposed here that for a complete understanding of speech perception, natural speech, vocoded speech, and synthetic speech must be evaluated in a variety of experimental tasks.

Chapter II

Literature Review

Models of speech perception are generally based on experiments conducted in unnatural, laboratory conditions. While this is an important first step, it is also important to include more realistic listening situations in any complete model. The speech signal itself may be generated (or re-generated) by a variety of **sources**. For example, speech may be spoken in a quiet background by a known individual with a similar dialect to the listener. Or, it may be created by a synthetic voice providing directions on a GPS. Or, it may be spoken by a native speaker of the listener's dialect with a speech disorder. Models of speech perception must be able to handle signals at extreme and intermediate levels of speech coding quality.

In the present work, the following speech sources were studied: Sentences spoken by native speakers of the listener's dialect of English (in noise); computer synthesized speech (as in GPSs or augmentative communication devices); and vocoded speech (as in cell phones).

Speech may arrive via a variety of communication **channels**. The channels may range from extremely clear to unintelligibly distorted. In this work, cell phone encoding was selected as one of the many possible channels that speech may be conducted through. In this case, cell phone speech was presented in a background of noise. Note that vocoding techniques may be considered as a quality of the speech signal or a characteristic of the

communication channel (Lam, Ay, Chan, Hui, & Lau, 1996; Shacham, Craighill, & Poggio, 1983).

The speech environment consists of non-acoustic attributes as well. For example, speech may be combined with visual information such as facial motion. It may also be combined with additional tasks such as driving, taking notes in a classroom, or working in a factory. These factors also have significant effects on listener's perception. From the opposite perspective, listening to speech can also impose cognitive load on the listener that will affect performance on simultaneous tasks. For example, listening to speech has been demonstrated to distract listeners from visual tasks such as the mental rotation of three dimensional objects (Johnsin, 19xx).

Therefore, the ability to perceive speech is based on external listening environments as well as the specific qualities of the speech signal itself. The internal processing brought to the task of perceiving speech is multi-dimensional and a wide range of the listener's cognitive capabilities must cooperate to accomplish the act of speech perception. What may appear to be a relatively simple task of decoding an acoustic pattern into a pattern of phonemes or words turns out to rest on many cooperating processes.

Taken together, cognitive load and attentional factors play a major role in constraining the ability to perform simultaneous tasks while listening to different types of speech in complex listening environments. To understand how listeners perceive speech under these demands one must begin by

examining the concepts that underlie the relationships between the acoustics, phonetics, and semantics of speech. For the present discussion the focus will be on theories of speech perception and how they might apply to different speech sources (e.g., natural, reverberant, or synthetic) while listeners perform multiple simultaneous tasks.

Theories of Speech Perception

Any complete theory of speech perception must address the following basic fact – There is no straightforward, one-to-one correspondence between a speech segment and its acoustic qualities (Pisoni, 1993). Researchers agree that speech perception does not simply involve the direct translation of acoustic signal to a semantic understanding of the information intended. This is called the problem of the "lack of acoustic-phonetic invariance". Prior knowledge about the stimuli presented, characteristics of the individual phonemes, and context appear to play important roles in the perception of speech. Several of the more popular theories that attempt to account for this issue are described below. To the extent that they currently perform well, these theories must now be extended to account for speech complex contexts. The goal of the present work is to provide data to help define the interaction between the traditional issues in speech perception and the more complex issues of speech perception that involve difficult communication channels, noise, and complex environments.

Bottom-up Processes in Speech Perception.

Bottom-up processing based models are characterized by information flow from the periphery to the central nervous system: information is observed in an acoustic waveform, combined to provide meaningful auditory cues, and passed to higher level processes for further encoding. One of the early theories that used bottom-up processing techniques to explain speech perception was the Lexical Access From Spectra (LAFS) (Klatt, 1978) model. The LAFS system is a computer algorithm for efficient accurate lexical search. It is based on a relatively simple spectral template matching system. Every word is represented with a stored sequence of power spectra. The input signal is converted to a similar form and the closest matches are identified as being the same word. However, LAFS provides no explicit way to learn new words or combine words in new ways. LAFS was developed to demonstrate how accurate word recognition could be with no top-down information. That is, it was conceived as a standard against which to measure the performance of top-down theories.

The problem of acoustic- phonetic non-invariance was addressed in LAFS by using spectral-sequence diphone definitions and recognition strategies. LAFS avoids explicit phonetic transcription by precompiling knowledge of acoustic-phonetic interactions into the diphones and lexical definitions. More specifically, the problem of representing words in the lexicon for optimal search was solved by converting abstract phonemic forms

into spectral sequences. Representing phonological recoding across word boundaries was solved by applying a set of phonological rules to augment and adjust the connectivity pattern of the lexical network. Recovering from wrong phonetic decisions was nullified by not making intermediate phonetic decisions. Interpretation of phonetic cues was solved by sorting expected prosodic attributes of words on the lexical decoding network and by interpreting deviations from the expectations.

Even though all these problems were addressed using the LAFS model, it was not able to address issues such as the perception, learning, and understanding of new words. It could also not cope with speech in noise findings, the McGurk effect or phonemic restoration. These issues have been found to be better addressed by models that used top-down information.

Top-down Processes in Speech Perception.

Top-down processing models assume internal, high-level information of the acoustic environment and prior knowledge of the properties and dependencies of the objects in it. In models that are strictly top-down, information flows from central to peripheral locations in the nervous systems. For example, the peripheral system is seen to collect evidence that would either justify or cause a change in a central world model and in the state of the objects in it. This approach is also called prediction driven processing, as it is strongly dependent on the predictions of an abstracted internal model, and on prior knowledge of the sound sources (Ellis, 1996). It should be noted

that most modern top-down processing does not require that most of the information must have originated at a high-level of processing. Top-down processing is highly interactive and able to adapt to the signal at hand. For example, context is utilized when information is collected at a low level, interpreted at a higher level, and brought back to affect low level processes. Some frequently replicated examples that support top-down processing are the *phoneme restoration effect* (perceiving an entire utterance as intact when a part of the utterance is completely removed and replaced by another sound such as a cough, a tone, or white noise; Warren, 1970), the *word superiority effect* (the demonstration that listeners perceive phonemes more rapidly and more accurately in a word as compared to perceiving the same phoneme in isolation; Reicher, 1969), and *sine wave speech perception* (perceiving an unnatural acoustic signal consisting of three or four time varying sinusoids as noise unless listeners are told in advance that the tones are speech; Carrell & Opie, 1992).

There exist listening situations where specific words can be anticipated on the basis of situational context and prior dialog. In such circumstances, the listener does not wait for wait for the bottom-up analysis, but instead, the higher level top-down processing capabilities can scan the input as it arrives so as to make an early lexical decision and prepare for the next word. For example, while speaking to a gate agent at an airport it takes very little acoustic information to hear the word "itinerary." This effect has been

demonstrated at the phoneme, syllable, word, and sentence levels (Samuel, 1981).

Theories of Speech Perception.

Theories of speech perception fall into two major classes: Those assuming articulation-based units of speech and those assuming acoustic-based units of speech.

Articulation based theories.

One popular theory of speech perception is the *Motor Theory of Speech Perception* (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). This theory postulates that speech is perceived by reference to how it is produced. According to motor theory, listeners access their own knowledge of how speech is articulated to decode speech into articulatory motion. Articulatory gestures such as pressing the lips together or rounding the lips are the fundamental units of perception that provide listener with linguistic and semantic information. According to motor theory, the ways that phonemes are produced and perceived have more in common than the ways they are acoustically represented and perceived. A number of studies have provided strong support for the motor theory. The McGurk Effect (Mcgurk & McDonald, 1976) demonstrated that visual articulatory information can just as powerful as acoustic information in phoneme, syllable, and word recognition. In the seminal work, it was found that when a visual / ga / is

presented with an auditory / ba /, the listeners perceived / da /. In addition, very young infants have been shown to look at pictures of faces with rounded lips when presented with the vowel / u / whereas they look at faces with spread lips when presented with the sound / i /. These results make it difficult to argue that articulation is not fundamental in some circumstances.

Acoustic-based theories.

The *cohort theory* is a qualitative description of word recognition (Marslen-Wilson, 1987, 1989; Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978). Cohort theory rests on the assumption that "each memory element in lexicon will be computationally active processing entity" (Marslen-Wilson & Welsh, 1978). In early versions of the model, these memory entities process both bottom-up acoustic-phonetic information and top-down syntactic-semantic information. Lexical access is only triggered by acoustic-phonetic input and thus no preselection of words based on context can occur. The initial, around 150 to 200 ms stretch of speech, however represented, activates in an all-or-none manner a cohort of items that match this initial representation. Activation makes available the phonological, syntactic, and semantic information associated with each lexical entry. As the acoustic-phonetic evidence accumulates over time, members of the cohort are affirmed and retained or disconfirmed and removed. Recognition occurs when the cohort is reduced to a single candidate. Although the cohort is driven by acoustic input, inappropriate syntactic and semantic information can

eliminate members of the cohort. It is at this intersection that bottom-up and top-down information are combined. Recognition is assumed to be optimally efficient in part because the processes of access and selection can make use of contextual information. Experiments using techniques that examined the time course of word recognition (Samuel, 1981) have supported the general claims of cohort theory.

The *TRACE* model of speech perception is also entirely acoustic but was derived from general neural network (connectionist) models, not speech-specific models. Trace is a quantitative model designed originally to recognize speech at the segmental level and later extended to simulate the recognition of words (Elman & McClelland, 1986; McClelland & Elman, 1986). TRACE is an interactive-activation model of processing in which information flows in both directions, bottom-up and top-down. The TRACE model consists of a network of large number of nodes, separated into three levels; feature, phoneme, and word. Each of these levels contains highly interconnected processing nodes. TRACE accounts for several different aspects of human speech perception. Like humans, TRACE uses information from overlapping portions of the speech wave to identify successive phonemes. These representations are processing elements (or units) that are activated in proportion to the strength of the hypothesis that a unit represents. The internal activation levels of the nodes are set by training rather than by pre-programming acoustic-phonetic information. Although there are variations in

different implementations, the training follows the same general strategy. The nodes all begin at random levels of activation. The system is then presented with pre-recorded speech signals repeatedly, billions of time, to its input. The correct utterance is provided to the output. When some portion of the output is corrected, the nodes with higher excitatory responses have their excitation level incremented further. Conversely, the nodes active during an incorrect trail have their excitation levels decremented. After the network is trained, it is relatively successful at recognizing speech that is new to it.

The *Neighborhood Activation Model* (NAM, Luce & Pisoni, 1998; Luce, Pisoni, & Goldinger, 1990) is a classic example of interactive activation and competition model of speech perception in which every stimulus input activates a set of similar sounding acoustic-phonetic patterns in memory. Once the patterns are activated, word decision units are tuned to decide the best matching pattern. The word decision units compute probabilities of each pattern in memory based on the frequency of the word to which the pattern corresponds, the activation level of the pattern, and the activation levels and frequencies of all other patterns activated in the system. The word decision unit with the highest probability is perceived. According to NAM, spoken words with many similar sounding neighbors would be processed more slowly and less accurately than words in low-density neighborhoods. Figure 1 shows a representation of NAM.

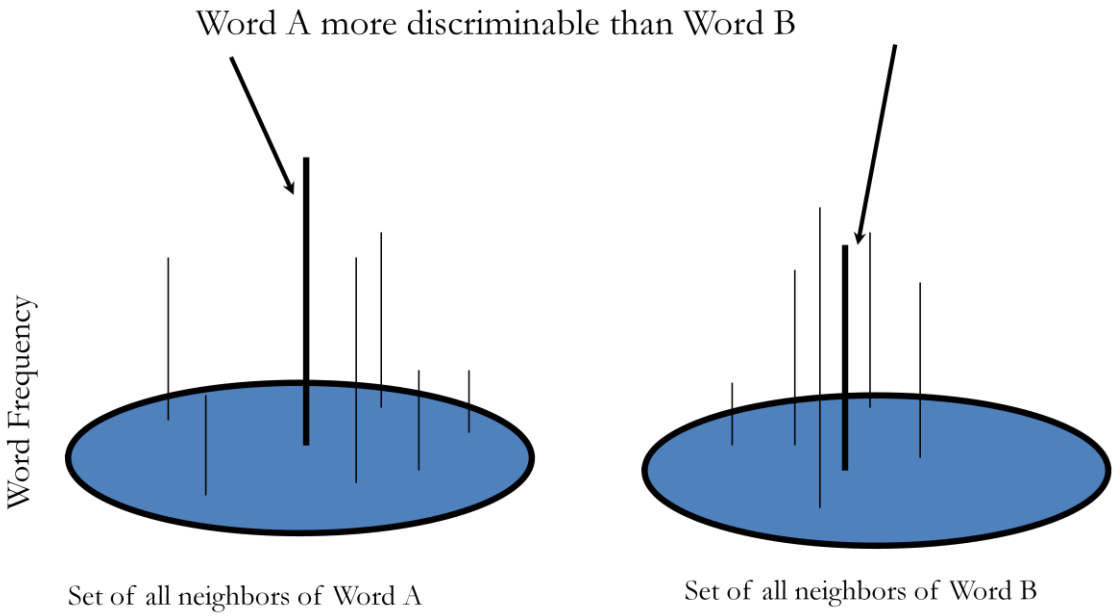

Figure 1. Neighborhood activation model

PARSYN (Luce, Goldinger, Auer, & Vitevitch, 2000) is a refinement of the NAM. PARSYN stands for PARadigmatic and SYNtagmatic referring to neighborhood activation and phonotactic constraint, respectively. PARSYN is a connectionist (neural network) model in which the networks change over time depending on their input. PARSYN is designed to simulate the effects of both neighborhood activation and the probabilistic phonotactics on the processing of spoken words. Effects of neighborhood density arise from the lateral inhibition at the lexical level. Effects of probabilistic phonotactics arise from activation levels of and interconnections among units at the allophone pattern level. PARSYN takes the mathematical skeleton of the NAM and provides a connectionist implementation. Allophones that occur more frequently have higher resting activation levels. Also, allophones that

frequently occur together will excite one another via facilitative links. This model resembles TRACE at the word level, but representations at the input and intermediate layers are allophones rather than features or phonemes. PARSYN was developed to overcome two significant shortcomings of NAM. PARSYN better accounts for temporal dynamics of recognition, including the ebb and flow of neighborhood activation and phonotactic constraints. Also, PARSYN incorporates a sub-lexical (morphonemic) level of representation missing in the original NAM. However, it should be noted that the ability of the PARSYN network to accomplish self-training is the major reason for its performance difference with NAM.

Limitations of current theories of speech perception.

Theories of speech perception often take talker differences such as age, gender, and dialect into account. Coarticulation and speaking rate are also embedded into many models (such as the theories described above). However most of the models of speech perception begin with a clear signal, free from reverberation, distortion, and noise. They are primarily based on data collected in the laboratory in which the listener has only one task. Because of this, most theories ignore two important problems encountered in real-world communication: 1) Speech is often embedded in environmental noise and distorted in a variety of ways (Assman and Summerfield, 2004; Teder, 1990). 2) Speech is often produced and perceived during the simultaneous performance of unrelated tasks. In everyday situations we rarely devote

complete attention to what we are listening to. While the neglect of noise and the attention demands of multiple tasks have always been problematic for models of speech perception, these issues have become more prevalent as technology makes listener's environments more complex.

The present work investigated the perception of vocoded and synthetic speech as listeners performed simultaneous visual-motor tasks. This combination of tasks was selected for study because it is important for understanding speech perception ion typical listening environments and because it is generalizable to a variety of more complex communication situations. For example, this situation is analogous to cell phone listeners driving cars or trucks, piloting aircraft, or controlling construction or industrial process. Very little is known about the impact of incomplete speech cues on attention and performance with simultaneous tasks. Conversely, little is known about the effect of simultaneous visual-motor tasks on speech intelligibility.

Although theoretical frameworks in speech perception take coarticulation, speaking rate, talker idiolects, and many other internal sources of variation into, they typically ignore the problem of extracting speech from concurrent acoustic input. The process of speech perception involves organizing *acoustic information* into meaningful units (Jusczyk, 1997) yet most studies have only worked on organizing *speech information* into meaningful units, not the entire sound field. It is only after general

perceptual organization has occurred that phonetic and linguistic meaning can be assigned. Widely accepted theories of word recognition (e.g. TRACE (McClelland & Elman, 1986) and PARSYN (Luce, Goldinger, Auer, & Vitevitch, 2000) have addressed the acoustic organization of speech but they have not addressed how listeners organize acoustic information *prior* to the activation of acoustic-phonetic and word recognition processes. That is, they do not address how speech is separated from simultaneous sounds (Bregman, 1990). The basic sounds upon which speech perception theories are founded are almost never presented in isolation to listeners. Instead, current theories are typically defined with the assumption that the only sounds present are speech sounds. Not surprisingly, these theories are often developed using experiments in which the acoustic signal is presented without reverberation, distortion or delay. To the extent that additional sounds are considered, they are typically mathematically tractable signals such as white noise, speech-shaped noise, or band-pass filtered speech. Unfortunately, environments outside of laboratory settings rarely deliver signals with these helpful characteristics.

The extent of this deficit in our knowledge is well-illustrated by the difficulty that computer speech recognition systems have with recognizing human speech (Fucci, Reynolds, Bettagere, & Gonzales, 1995; Lai, Wood, & Considine, 2000; Pisoni, Nusbaum & Greene, 1985; Reynolds, Bond, & Fucci, 1996; Scherz & Beer, 1995; Venkatagiri, 2003). These programs often do

exceptionally well when recognizing speech in extremely quiet conditions, for example, when close talking microphones are used in quiet, non-reverberant conditions. However, automatic speech recognition fails when performing in more realistic situations even those in which humans are able to understand speech with little difficulty. For example, a typical student sitting in the center of a quiet classroom will have almost no difficulty hearing the instructor whereas a speech recognition system placed in the same location as that student will fail spectacularly. To the extent that speech recognition systems instantiate our theories of speech perception, the failure of automatic speech recognition in natural environments demonstrates large gaps in our current theories (if not entire failure).

Description of Signals to be Investigated

In natural environments, individuals are presented with a large variety of distorted, incomplete, and poorly represented speech signals. Despite a strong desire to investigate a wide array of difficult stimuli in order to discover general classes, rules, and limitations of these varieties of speech, only a few were selected for the present work due to resource and time constraints. The difficult speech sources that are tested in experiments 1, 2, and 3 are discussed in detail below. This will allow the precise differences between the acoustic construction of these stimuli and that of more-typically-studied stimuli to be described. In addition, several types of speech that will

not be investigated here are also described in order to provide a more complete context for the stimuli that were studied.

Telephony.

Emergence of telephony on cell phones and the internet has increased the range of environments in which humans commonly communicate (Lam, Ay, Chan, Hui, & Lau, 1996; Markopaulou, Tobagi & Karam, 2002; Narbutt, Kelly, Murphy, & Perry, 2005; Tan, Wänstedt, & Heikkilä., 2001). Even though advanced telephony techniques are cheap and reliable, they often exhibit jitter, echo, loss of data packets, delay and low-bitrate vocoding. Because of the role of technology in our lives, a description of speech perception is not complete without a discussion of modern telephony techniques which include circuit switching networks, packet switching networks and voice over internet protocols. For example, speech transmitted via a cell phone network is severely degraded due to the vocoding algorithms, noise in the channel, and delay. This is a very different signal than quiet rooms or reverberant, noisy offices.

Traditional telephone systems are based on *circuit switching networks.* These networks establish a fixed bandwidth path between nodes and terminals before the users may start communicate. It is as if the nodes were physically connected with a single electrical circuit (Martin, 1976; Spragins, Hammond & Pawlikowski, 1991; Tanenbaum, 1996). The Public Switch Telephone Network (PSTN) is an example of a circuit switching network.

Each connected circuit cannot be used by other callers until the circuit is released and a new connection is set up. Even if actual communication is not taking place in the dedicated network, the channel remains unavailable to other users. Due to the existence of a dedicated network between the source and the destination, the entire message is sent in order. Circuit switching is relatively inefficient because channel capacity is wasted on connections which are set up but not in continuous use. On the other hand, the connection is immediately available to a subscriber and the capacity is guaranteed until the call is disconnected. Analog speech representation and less compression make the traditional telephone connection more similar to live speech than more modern techniques.

More modern telephone systems use p*acket switching*. These networks segment a message/data to be transmitted into smaller packets between about ten to twenty milliseconds long (although this varies tremendously based on the specific network). The packets can take any available path from source to reach the destination (Kurose & Ross, 2004; Martin, 1976; Rappaport, 1996; Spragins, Hammond & Pawlikowski, 1991; Tanenbaum, 1996; Turner, 1988). Each packet is labeled with its destination and packet number removing the necessity of a dedicated path from the source to destination for the packets to reach. The routing of the packets is done using a routing algorithm. The routing algorithm can create paths based on various metrics and desirable qualities of the routing path. It is also entirely possible

that two successive packets can take totally different routes to reach the destination. Once the packets reach the destination, the message is reassembled using the packet number. If some packets are lost during transmission, the receiver can request the lost packets to complete the message. On the other hand, the receiver can predict the information in the lost packets based on previous samples. Packet switching is used to optimize the channel capacity available in a network, to minimize the transmission latency and increase the robustness of the communication system.

Voice over Internet Protocol (VoIP) is a protocol optimized for the transmission of voice over internet and other packet switched networks (Breslau et al., 2000; Ding et al, 2007; Douskalis, 2000; Wang, Liew & Li, 2005). VoIP uses a combination of packet switching and digital speech compression to provide very efficient speech communication. Cost savings can be achieved by using VoIP to transmit audio especially where users have existing network which is underutilized. Because VoIP uses an internet connection, it is susceptible to all the disturbances associated with general-purpose broadband services in addition to disturbances that are uniquely problematic for voice communication. Some of the problems which are apparent on a VoIP speech transfer are latency, jitter and packet loss. In addition, significant delay between the speech being spoken by the speaker and the listener can interfere with word and sentence recognition as well as conversational pragmatics. In addition to data transfer issues, VoIP signals

are typically vocoded to save bandwidth costs and this coding is imperfect. The marketplace is used to determine what level of coding representation is "good enough".

Cell phone speech is also coded in a similar fashion to VoIP signals. Although the specific codecs (coders/decoders) are viewed as highly proprietary intellectual property by their owners, it is clear that they use a compression algorithm generally named "*Code Excited Linear Prediction*" (CELP). The compressed data is transmitted over the network using packet switching technology (Schroeder & Atal, 1985).

The linear prediction system models (to a large degree) the structure and function of the vocal tract. This is used to constrain the sounds that must be transmitted so that they are only those that a human can produce. This greatly reduces the information that must be transferred because, for example, earlier occurring sounds may be used to predict the following ones. To a lesser degree, knowledge of psychoacoustics is also used to transmit only data that is relevant to the human auditory system. Psychoacoustic information is implemented in vocoding algorithms by reducing the number of bits used to quantify basic acoustic parameters to noticeable levels, but are considered to have minimal effects on intelligibility.

Two additional issues arise when investigating cell phone speech. When speech is highly encoded, not only is intelligibility is degraded but non-linguistic information such as speaker identity, emotions, sincerity,

intonation, timbre, age, and size information are degraded as well. Therefore, it is possible that highly encoded speech may be intelligible, yet annoying or distracting to the listener. The second important consideration when making claims about cell phone speech is that most of the encoding algorithms are proprietary. This makes it difficult to make generalizations about cell phone signals. However, it was felt that, because they are based on CELP, the algorithms are similar enough that meaningful results might be discovered from investigating just one vocoding system. If sufficiently compelling, these results could lead to work with a wider variety of systems.

Because cell phones are ubiquitous and will form a major part of the stimulus material, the representation of speech in the cell phone network and handsets will be discussed in detail here. The CELP algorithm is based on five main premises:

- The source-filter theory of speech production as implemented via a linear prediction model affords a sufficient representation of speech
- It is possible to discard much of the speaker's voicing source and transmit a very low bit-rate voicing source. This may be accomplished via the use of an adaptive and a fixed codebook which approximate "average" glottal sources. This is used as the excitation signal for the linear prediction model to reduce bandwidth requirements. However, it also makes all voices sound more similar to one another.

- Search is used in a "perceptually weighted domain" to find the best representation of the signal that will be available to the receiver using knowledge of listener's auditory system.
- Vector quantization is used extensively. It is an algorithm for reducing the information to the minimum number of bits necessary to convey speech for any given parameter (e.g., fundamental, pole 1, etc.). These numbers can be increased and decreased as bandwidth availability and radio-frequency interference allow more and less information to be sent.
- An analysis-by-synthesis technique is used to decode the signal. That is, a number of different decodings are created. These versions are then re-synthesized and the one that is analyzed (by the receiving instrument) to be the best sequence of sounds is selected to be the output.

Source-filter model of speech production.

The source-filter model of speech production is based on the assumption that the vocal folds are the source of a harmonic sequence (the excitation signal) which drops in amplitude between 8 and 12 dB per octave. The vocal tract acts as a filter to spectrally shape the source and thereby produces the various sounds of speech. This is an approximate model of speech production and is widely used for its simplicity and easy implementation. The different phonemes produced can be distinguished by

their source and filter shape. Voiced sounds (vowels) have an excitation signal that is periodic. In source-filter theory, the periodicity is approximated by an impulse train in the time domain. Voiceless fricatives (such as / s /, / θ /, and / f /) have an excitation signal similar to white Gaussian noise. The voiced fricatives (such as / z /, / v /, and / ð /) have excitation signal composed of a harmonic part and a noisy part. Filters are created by the positions of the articulators and shape the source spectrum in the frequency domain.

Linear prediction.

The source-filter mechanism is often modeled by linear predictive coding (LPC). Linear prediction is the basis of CELP. Linear prediction predicts the signal $x[n]$ using a linear combination of its previous samples:

$$y[n] = \sum_{i=1}^{N} a_i x[n-i]$$

where $y[n]$ is the linear prediction of $x[n]$. The prediction error is the error between the original signal and the predicted signal and is given by:

$$e[n] = x[n] - y[n] = x[n] - \sum_{i=1}^{N} a_i x[n-i]$$

This works because each point in a speech waveform is reasonably predictable given several dozens of previous points. The goal of LPC analysis is to find the best prediction coefficients a_i which minimize the quadratic error function:

$$E = \sum_{n=0}^{L-1} [e[n]]^2 = \sum \left[x[n] - \sum_{i=1}^{N} a_i x[n-i] \right]^2$$

A fortuitous benefit of calculating these values is that speech formants are directly derived by the LPC coefficients.

Pitch prediction.

During voiced segments, the speech signal is periodic. Hence, the excitation signal $e[n]$ is approximately modeled as a factor of pitch period and pitch gain.

$$e[n] \cong p[n] = \beta e[n-T]$$

where T is the pitch period, β is the pitch gain.

Innovation codebook.

The final excitation $e[n]$ is the sum of the pitch prediction and an innovation signal $c[n]$ taken from a fixed codebook. The innovation signal is also called the "error term" as represents the part of the signal that was not predictable. The final excitation is given by

$$e[n] = p[n] + c[n] = \beta e[n-T] + c[n]$$

Noise weighting.

Modern audio codecs "shape" the noise so that it appears in the frequency regions where the ear cannot detect it as the human ear is more tolerant to noise in parts of the spectrum that are louder. CELP minimizes

the mean square error of the noise signal so that the signal could be properly reconstructed.

Vector quantization.

Vector quantization is a classic technique which reduces the amount of information that is necessary to store the signal or to transmit it. It works by dividing a large set of points into groups having approximately the same number of points closest to them. Each group is represented by its centroid point obtained by clustering algorithms. The density matching property of vector quantization is powerful especially for identifying the density of large and high-dimensional data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. This feature makes vector quantization for lossy data compression and data transmission practical.

Analysis-by-Synthesis.

Analysis-by-Synthesis (AbS) is the process of re-synthesizing the speech signal based on the transmitted parameters. Several signals are produced by optimizing the received parameters and the best sounding signal is selected as the transmitted signal. The optimization is done using small, sequential search algorithms to overcome the problem of limited computing resources.

CELP vocoding results in an extremely low-bitrate representation of speech. This comes at the expense of intelligibility and naturalness as perceived by listeners. However, it does allow an extremely efficient use of bandwidth which has driven great advances in technology. Also, CELP introduces delay in the system that is more noticeable when the networks are congested and processing power is low. Vocoding systems have been optimized to use only the minimum bandwidth required to make the caller's voice recognizable to the receiver and to recreate most of the sounds spoken so that the speech is intelligible. If only intelligibility and naturalness of speech representation are considered then the quality versus expense of speech tradeoffs are probably best determined by a marketplace to determine. However, if the implementation of these algorithms results in poor performance during secondary tasks it is worth further investigation to determine precisely what the tradeoffs are.

Recall ability for synthetic and natural monosyllabic words presented at different presentation levels were compared and found that free recall was consistently poorer for synthetic lists and the decrement did not increase differentially at faster presentation rates (Luce, Feustel, & Pisoni, 1983). Also, differences in ordered recall between the synthetic and natural word lists were substantially larger for the primacy portion of the serial position curve than the recency portion. They concluded that the difficulties observed in the perception and comprehension of synthetic speech was due to the

increased processing demands in short term memory. They suggested that the greater difficulties experienced with the perception and comprehension of information presented in synthetic speech, relative to natural speech, have three possible causes namely prosodic inadequacies, inadequate specification of acoustic cues to phonetic segments, and more processing needed to maintain information in short-term memory. These three causes may not be necessarily mutually exclusive. Nooteboom (1983) found that the recall of synthetically presented sentences is enhanced by the insertion of grammatical pauses whereas for natural speech, it is not.

Synthetic speech is less intelligible in the presence of background noise at a 10 dB signal-to-noise ratio than in ideal, quiet listening conditions (Fucci, Reynolds, Bettagere, & Gonzales, 1995). Also, listeners made different types of segmental errors with synthetic syllables in the presence of background noise than they did with natural speech syllables (Nusbaum et al, 1984). This was specifically true for DECTalk synthetic speech, even at a signal-to-noise ratio as high as 28 dB. Pisoni, Nusbaum, & Greene (1985) attributed the significant degradation in the intelligibility of synthetic speech in the presence of background noise to the absence of redundant segmental cues which were present in the natural speech.

Channel

One of the basic problems in the study of speech perception is how to deal with the inadvertently added noise to the speech signal. This is shown

by the pattern of difficulties that computer speech recognition systems have with recognizing human speech. These programs can do well at recognizing speech when they have been trained under quiet conditions. However, these systems do poorly with more realistic listening channels where humans are able to understand speech without difficulty.

Effect of noise.

Noise was one of the critical factors that affected a person's speech intelligibility (Hawley, 1977). Noise interferes with speech in various ways. Increase in noise lowers the intelligibility of speech. Also, increase in background noise changes individual's hearing threshold and affects the individual's ability to listen to speech.

To improve the ecological validity of speech perception tests, it has been argued that the masking noise needs to be similar to that found in listener's everyday listening environment (Sperry, Wiley, & Chial, 1977). In order for speech perception tests to contain improved ecological validity, they should be conducted in the presence of background noise (Larsby & Arlinger, 1994). Listeners with or without hearing impairment have difficulty understanding speech and communication in situations where the level of background noise is sufficient to mask speech (Dirks, Morgan, & Dubno, 1982; Helfer, 1991). The more similar the spectral characteristics of the masking noise to the speech signal (speech-shaped noise versus narrow band noise) the greater the masking effect (Dirks & Bower, 1969; Larsby &

Arlinger, 1994; Miller, 1947). Conversely, if the maskers are wideband and uniform, or narrow and variant, its masking effect is smaller than that of speech-shaped noise (Sperry, Wiley, & Chial, 1997).

Low redundancy sentences each were used to determine the masking effect of amplitude modulated noise and steady-state noise on perception of speech (Gustaffson & Arlinger, 1991). In the experiment, listeners in normal-hearing group were presented stimuli at 65 dB SPL. For the hearing impaired group, the stimuli were presented at a comfortable SPL chosen by the listeners. They reported that amplitude modulation improved speech perception for both the groups. The authors concluded that the participants obtained a release of speech masking when the noise was amplitude modulated; however, the effect of masking release was opposite with the unmodulated noise.

Time Varying Sinusoids (TVS) speech was used to determine the contribution of amplitude comodulation to auditory grouping in speech perception (Carrel and Opie, 1992). They compared recognition scores for unmodulated TVS (UTVS) sentences to sentences where the three sinusoids were amplitude modulated simultaneously (AMTVS) at 100 Hz. Results revealed greatly improved intelligibility for the AMTVS sentences, supporting the hypothesis that the modulation served as mechanics for grouping the three sinusoids as one auditory object, thereby increasing intelligibility.

Effect of reverberation.

As speech travels from the mouth of the talker, the acoustical energy is spread over an increasingly large area and the average decibel level falls. To a first approximation, this effect follows the 6dB rule. That is, the average speech level falls by 6 dB for every doubling of distance from the lips. For example, if the average level is 70 dB SPL at 1 m from the lips, then it is 64 dB SPL at 2 m, 58 dB SPL at 3 m, and so on. In open air, listeners receive only the direct speech signal.

In enclosed spaces, listeners also receive speech via reverberation. Reverberation refers to the persistence of sound in a room because of multiple, repeated reflections from the boundaries. During sound segregation, the reverberant sound is more or less uniformly distributed throughout the room. The level of this reverberant sound in relation to the level of the original source depends on the room size, the absorptive properties of its boundaries, and the directionality of the sources (Davies & Davies, 1997). When the sound source stops, the reverberant sound level begins to fall but it takes some time for it to become inaudible. The time taken for the level of the reverberated sound to fall by 60 dB is called the reverberation time. This quantity provides a rough measure of the reverberant properties of a room. Reverberation times in large, reflective spaces such as gymnasia can be as high as 2 or 3 seconds. However, in small classrooms with many absorbent surfaces, reverberation times may be as low

as 0.3 or 0.4 seconds. At any point in the room, a listener receives both direct sound, whose level follows the 6 dB rule, and reverberant sound, whose level is relatively independent of the distance. When the listener is close to the source, the level of the direct sound exceeds that of the reverberant sound. When the listener is far from the source, the reverberant sound dominates.

In addition to background noise, it has been recommended that researchers include reverberation in tests of speech perception to simulate every day listening environments (Hawkins & Yacullo, 1984; Nabelek & Robinette, 1978). Studies evaluating the effects of reverberation showed that an increase in reverberation time distorted the speech signal, resulting in a significant decrease in speech intelligibility (Helfer, 1991; Nabelek & Mason, 1981; Nabelek & Robinette, 1978; Yacullo & Hawkins, 1987). People with normal hearing show some decline in speech intelligibility, but this decline is not as large as that received by people with sensorineural hearing loss (Nabelek & Pickett, 1974). The explanation may be that people with a hearing impairment are less capable of integrating the direct and reverberant sounds. Tests conducted with hearing impairment show a decrease in the perception of vowels (Nabelek & Letowski, 1985), consonants (Helfer & Wilber, 1990), words with carrier phrases, and sentences (Nabelek & Robinette, 1978; Nabelek & Robinson, 1982, Yacullo & Hawkins, 1987). The commonality is that despite differences in design and population tested,

all studies report an increase in reverberation time leads to a decrease in speech intelligibility.

Environment

Speech perception and vision.

Ideally, optimal communication occurs when the listener can both hear and see the speaker (Lindbloom, 1990; Sonnenschein, 1985). Although, both auditory and visual cues are available a good percentage of the time, it is not always the case. Still, speech perception is considered to be primarily an auditory process only. Vision helps to direct our attention to the speech signal and differentiate it from the surrounding auditory background noise, providing redundant and complimentary cues in addition to the auditory cues given (Boothroyd, 1982; Chen & Rao, 1998). All of this serves to improve our speech perception abilities. Therefore, even though we are not consciously aware of it, the combined auditory-visual systems are involved in communication in many situations.

The visual system aids localization by the additional input it provides. Visual input helps listeners to focus attention on the acoustic source, making it easier to localize due to the redundancy of the dual inputs. Also, dual inputs help in better understanding speech as is evident from higher intelligibility scores for normal talkers in environments such as class rooms,

office, and laboratory. When stimuli presented through auditory system is degraded, the information provided by the visual system can become critical to normal functioning in everyday life. When listeners talk over telephones and via other medium where they do not see the speaker, they are deprived of these additional visual cues. This reduction in cues makes the speech less intelligible which is evident from lower intelligibility scores especially in synthetic and vocoded speech. The negative effects of vision on speech perception will be considered later.

Attention and Speech Perception.

Attention is the mechanism by which certain aspects of the surrounding environment are selected for further processing while others are inhibited and it appears to be an integral part of most cognitive tasks (Hoffman, Yang, Bovaird, & Embretson, 2006). Research in many domains has demonstrated that attention has a limited capacity. Humans cannot attend to everything perceived using optical and auditory system. The human brain allocates resources to different simultaneous tasks according to the demands of the task.

It was argued that Broadbent's Filter theory (1958) and its revision by Treisman does not apply to tasks that required divided or selected attention (Moray, 1967). These models failed to explain the good performance of well-trained observers on tasks in which the inputs and outputs are compatible. In other words, the humans should be considered as limited capacity processor

instead of limited channel processor for better explaining results obtained in divided attention experiments. It is this limitation of the human processor that influences performance of different processes simultaneously that requires divided attention. These simultaneous mental operations were not considered in the theories by Broadbent and Treisman.

The single resource theory argued that there existed single limited pool of capacity was available for a variety of tasks and performance on these varieties of tasks depends upon the amount of resources allocated to the task (Kahneman, 1973). The overall capacity was a function of arousal and was limited. The participant devises an allocation strategy for dividing the available limited resources for various tasks based on the characteristics of the stimuli and individual motivation.

Multiple resource theory was an alternative to the single resource view of human capacity. According to Multiple resource theory, humans are multiple channel processors capable of doing several processes together (Navon & Gopher, 1979). The individual processors have their own limitations and capacities and are independent of other processors. Also, several processors could share the same available capacity. Because of the simultaneous existence of several individual processors, the effect of changing task difficulty on performance was multi-dimensional. Different combinations

of task difficulty and stimuli presentation may differentially affect the allocation of resources to individual processes.

According to Wickens' Multiple Resource theory (MRT) (Wickens, 1984), humans have several different pools of resources that could be tapped simultaneously while performing simultaneous tasks. The nature and difficulty of the simultaneous task dictated whether the resources were drawn from the same pools of resources or different pools of resources and whether the tasks were completed in sequential or in parallel. The tasks were performed in sequential manner if the tasks require accessing the same pool of resources. However, the tasks were performed in parallel if the tasks require accessing different pool of resources.

Wicken's model of resource allocation that assumes separate pools of resources for spatial and verbal information processing codes may be more appropriate for the study of speech perception and simultaneous visual-motor task as compared to the single resource model of attention. Listeners use efficient switching and scheduling while performing two or more activities over a short period of time (Wickens, 2000). Under extensive workload, human listeners may exhibit delayed information processing or even not respond to all of incoming information. This might happen due to the fact that the amount of information supplied surpasses the maximum capacity of information which human brain can process at a single instance of time.

When the listeners were forced to engage in concurrent processing, three further processes influence the effectiveness of multi-task performance. The three factors are confusion of task elements, cooperation between task processes, and competition for task resources. In contrast, when their mental workload is much lower, they may become bored and make mistakes. In this respect, human reliability has been defined as a function of the mental workload assigned (Rolfe & Lindsay, 1973).

Most physiological measures have been based on the single resource model of workload which stipulates that each individual has a limited processing capacity, with the cognitive mechanisms required to perform tasks and mental activities were viewed as a single pool of resources (Moray, 1967). This capacity could be allocated in graded amounts to various activities depending on their difficulty or demand for resources. In this respect, all task and mental activities shared the same pool of resources. As task demand increase, the central nervous system increases the supply to the pool of resources so that the task gets continued without any hindrance. Physiological measures are based on the concept that this general capacity or its manifestations can be measured.

The relation between resource allocation and task performance was supposed to be linear, until the moment all the resources were invested. From that point on, no more resources can be invested and task performance would remain stable. Norman & Bobrow (1975) called such a task resource-

limited which was opposite to data-limited task. When performing a data-limited task, additional available resource investment did not lead to increased performance due to the limitations in data quality. Although this theory could be applied to a variety of situations, it could not explain why effective time-sharing and unaffected performance could occur when a second auditory task was added to a primary task.

However, many researchers have shown that single resource theory may not provide adequate information to assess the workload associated with a multi-task operation. For example, the P300 amplitude was not found to be sensitive to increase in the difficulty of the tracking task when the number of number of tracked dimensions was increased (Wickens, Isreal, & Donchin, 1997). The fact that the P300 to the tracking task did not vary significantly as a function of the task difficulty was attributed to the idea that different tasks drew resources from different resource pools. This view asserted that although the tracking task difficulty tapped response-related resource, the counting task difficulty tapped perceptual-related resources. In similar studies, the P300 latency had been found to change with stimulus parameters such as masking that are known to affect encoding and central processing, but not for stimulus response processing (McCarthy & Donchin, 1981; Parasuraman, 1990). These results had been discussed in terms of the multiple resource model of workload. This model argues that several separate resource pools, instead of a single resource pool, exist corresponding to

different modalities, codes, and information processing stages (Wickens & Hollands, 2000).

Cognitive efforts in hearing.

When more than one message arrive at a particular instant of time, the capacity of the listeners to process all the messages were limited. However, when the listeners were instructed to focus on one particular message in the group of all the messaged, the listeners could selectively listen to that particular message and perform better on the selective listening task (Broadbent, 1952). The results of these experiments could not be explained purely based on sensory phenomena. Based on the results of selective listening experiments, it has been generally agreed that, to some extent when more than one stimulus is presented at any point of time, the stimuli may be dealt simultaneously provided they convey little information. The overall conclusion is that the listener can process only a certain amount of information at any given point of time (Broadbent, 1958).

The importance of cognitive processes on listening was also found in synthetic speech research. The acoustic characteristics of synthetic speech are very different from that of natural speech. The naturalness of sound measured in subjective tests was also frequently reported along with the intelligibility scores. It can be found that, in some cases, the synthetic speech may have the same intelligibility as natural speech; however, they differ substantially in terms of naturalness.

It was assumed the listeners are information processors with limited capabilities but are extremely flexible in accessing higher levels of information in sophisticated ways (Pisoni, 1982). Evidence from synthetic speech studies suggested that listeners perceive speech by using active top-down processes. Listeners can generally perceive speech to a greater extent under conditions with ambient noise, reverberation, or information overloading, as may exist in a degraded signal such as synthetic speech. Hence it can be concluded that a listener's actual performance cannot be precisely quantified based on the intelligibility scores measure from a single listening test.

Driving is one of the most important and interesting topics currently related to attention and speech perception. Speech communication and driving carries a wider range of theoretical and practical applications.

Cognitive Processing

Context.

The accuracy of the speech recognition not only depends on the sensory data generated from the stimulus itself but also from the context within which the stimulus occurs. For example, words are recognized more easily when they are presented in sentences rather than in isolation or in carrier phrases. This phenomenon was first investigated by Miller, Heise, & Lichten (1951). They found that at some signal-to-noise ratios, accuracy of recognition

of words increased as much as 20 percent when sentence context was added. Also, it has been shown that the effect of sentence context can be controlled by changing the extent to which the specific context limits the number of semantically plausible alternatives (Kalikow, Stevens, & Elliott, 1977).

An example of a context effect is the improvement of word recognition in sentences that occurs when the listener is given prior knowledge of sentence topic which generally exists in most conversations. A more subtle kind of context is that provided by the properties of the lexicon from which the words are drawn. It has been shown that real words, presented in isolation, are more easily recognized than are nonsense syllables and that words with a high frequency of occurrence in spoken and written language are more easily recognized than are words with a low frequency of occurrence (Giolas and Epstein, 1963; Savin,1963; Schultz, 1964).

The common factor governing the effect of context on speech perception is that context influences the apriori probability of the stimulus pattern being presented. When the apriori probability increases, the probability of correct recognition also increases. Boothroyd and Nittrouer (1988) measured percent recognition of phonemes and whole syllables in consonant-vowel-consonant (CVC) words and CVC nonsense syllables at different signal-to-noise ratios (SNR). They found that lexical, syntactic and semantic constraints served to increase the recognition probabilities for phonemes and words presented in noise.

Types of cognitive processing.

Two different types of cognitive processing, controlled and automatic, have been studied for over a century (James, 1890). A distinction was drawn between primary and secondary memory. Primary memory traces were active representations in memory, subjected to degrading through interference. Secondary memory was conceived as the more permanent repository of experience. Primary memory was argued to contain a kind of real-time record of temporal order information and analysis of its traces was assumed to form the basis for performance in immediate memory tasks. Also, memory traces were represented in primary and secondary memory as vectors or lists of features that can differ in value and type (Bower, 1967; Eich, 1982).

A more refined notion of dual processing has been prominent in the work of Richard Shiffrin and colleagues for the past 30 years. Specifically, the role of controlled versus automatic processing in studies of short-term memory and verbal learning was extensively examined. It is central to the theory of short-term memory (STM) and long-term memory (LTM) proposed by Atkinson and Shiffrin (1968). According to their theoretical framework, information from the environment arrives into a temporary short-term storage which served as an antechamber to the more durable LTM. In their model, the temporary system also served as a working memory, a workspace necessary not only for long-term learning, but also for many other complex activities such as comprehension and reasoning.

Automatic processing.

In Schneider and Shiffrin (1977), an automatic process was defined as the activation of a sequence of nodes that "nearly always becomes active in response to a particular input configuration," and that "is activated automatically without the necessity for active control or attention by the subject". The ability for the process to occur in the absence of control and proper attention by the individual was the basis for referring to such processes as "automatic." An automatic attention response is a special type of process that directs attention automatically to a target stimulus (Schneider & Shiffrin, 1977). Automatic processing is a fast, parallel, fairly effortless process that is not limited by short-term memory (STM) capacity, is not under direct subject control, and is responsible for the performance of well-developed skilled behaviors. Automatic processes are unintentional and unconscious and therefore they are not subject to conscious control, cannot be avoided, and cannot be terminated in the mid-operation. Automatic processing is the result of extensive training on exactly the same task (Schneider & Fisk, 1982). In the laboratory, this corresponds to many thousands of trials. Thus, it activates nodes in memory, but does not modify long-term memory (Fisk & Schneider, 1984). A classic example of automatic processing would be to stop the car at a signal when the light is red.

The contrast between automatic and controlled processes was initially studied using extended consistent mapping (CM) training (Schneider &

Shiffrin, 1977). During consistent mapping tasks, the participants' response to the stimulus presented is consistent across extended period of time. A search task becomes a CM task when the set of target stimuli is constant throughout the experiment. CM occurs when targets (the items being searched for) were never distractors (the items to be ignored) and distractors were never targets. In CM condition, the target and the distractor stimuli were chosen from disjoint sets. Under consistent mapping, automatic processes can develop slowly as repeated stimuli are attended to. Full development of automaticity requires thousands of trials even though performance improvements can be seen within the first few trials (Logan, 1992). For CM training, the memory set and the distractor set were always different. If numbers formed the memory set, then alphabets formed the distractor set and vice versa. Increasing the memory or distractor sets had no effect on the task performance. For example, if an experiment uses 1 and 9 as targets and A, B, C, and D as distractors, then the participants are using CM technique to identify the targets.

Controlled processing.

In Schneider and Shiffrin (1977), a controlled process was defined as "a temporary sequence of nodes activated under control of, and through attention, by the subject". Controlled processes are tightly capacity limited and are easy to set up, alter, and apply in novel situations for automaticity could never have been learnt. Controlled processing is a slow, serial, fairly

effortful process that is limited by short-term (STM) memory capacity, and is responsible for the performance of under-developed behaviors. Also, it requires substantial effort and interferes with other controlled processing tasks. A classic example of controlled processing would be to look for an exit with a particular restaurant while driving on a new city. Controlled processing is generally studied using varied mapping techniques.

In varied mapping (VM) training, the relationship of the stimulus to response mapping varies from block to block. A search task becomes a VM task when the stimulus that is assigned a given response on one block is assigned a different response on the next block. With varied mapping, the prior and current associations are incompatible, thereby making it impossible for the development of an automatic attention response and automaticity. VM occurs when targets may be distractors and distractors may be targets. In VM training, the memory set and the distractor set were always the same. A particular letter can be in the memory set in one block and then switch to the distractor set in a later block. Increasing the memory or distractor sets require slower presentation rates and the performance decreases even for slower presentation rates. For example, if an experiment uses 1, 2, R, and T as targets and 1, 3, Y, and T as distractors, then the participants are using VM technique to identify the targets.

Automatic processing is used for skilled behaviors. It includes the detection of familiar stimuli and initiation of a proper response. Automatic

processing is the result of extensive training in exactly the same task (Schneider & Fisk, 1982). Thus, it activated nodes in memory but does not modify long-term memory (Fisk & Schneider, 1984). Controlled processing includes effortful attentional memory search, learning and decision making. Also, controlled processing is sensitive to task difficulty which limits dual task performance (Fisk & Schneider, 1983).

Performance

Pisoni (1982) argued that listener's overall performance in a given task or situation is constrained by three factors namely

1. Processing limitations of the cognitive system
2. Fidelity of the input speech signal
3. Specific task demands of the human observer

The first constraint, processing limitations of the cognitive system, occurs because of the fact that acoustic energy had to be integrated over time in order to process constantly changing speech information. However, humans do have processing limitations to perceive, encode and store in a short-term memory until the corresponding information is encoded by the long-term memory. The STM is severely limited by the listener's past experience with the stimulus, listener's attention and the quality of the sensory input provided.

The second constraint, fidelity of the input speech signal, occurs based on the structure of the speech signal and can be compensated by

comprehension. It should be kept in mind that intelligibility is different from comprehension. Intelligibility is normally defined as the performance in the levels of phoneme, syllables, word, and sentence. Comprehension enables the user to arrange the speech units into meaning using linguistic rules. It helps the listener in predicting arrangement of sounds based on phonological rules or predict words based on semantics and syntax.

The third constraint, specific task demands of the human observer, is based on the demands that different task enforce on the listener. Humans are capable of developing perceptual and cognitive strategies to maximize performance under different simultaneous task conditions. The different strategies are adopted depending on the task presented to the listener. Hence, the study of these strategies is important in evaluating the effectiveness of any speech perception model and speech processing devices.

Dual-task performance and listening effort.

The use of dual-task paradigms to measure listening effort due to increased processing demands is based upon theories of attention and capacity limitation. Humans have a limited capacity for processing information at any given instant of time and simultaneous tasks that require capacity to be performed interferes with one another.

The relationship between performance and effort was first discussed by Broadbent (1952, 1958) based upon his observation that similar speech intelligibility scores could be obtained under various conditions at the

expense of unequal amounts of resources allocated by the listeners. Also, a listener who could correctly report speech presented in a noisy background might be less competent if the listener is required to perform a simultaneous unrelated task. It was found that there was a decrement in the simultaneous tracking task when subjects were listening to frequency-transposed speech as opposed to filtered speech. The difference in the two kinds of stimuli was reflected in the secondary tracking task performance. However, there was no difference in performance in the primary listening task. Based on these results, it was concluded that intelligibility tests did not differentiate listening effort from overall intelligibility performance and different tests should be devised in order to clearly evaluate the effect of intelligibility without the component of effort (Broadbent, 1958). Broadbent also suggested the importance of using multiple testing criteria within the same experiment to precisely evaluate intelligibility measures.

Downs and Crum (1978) used a dual-task paradigm to demonstrate the effectiveness of dual-task paradigms in quantifying processing demands under adverse listening conditions. The primary task was to repeat words presented at different levels and the secondary task was to respond to randomly presented visual stimuli. There was no difference in word recognition performance due to the introduction of noise. However, there was a significant increase in the reaction time to respond to the visual stimulus as noise increased. The authors attributed this increase in reaction time to the

increase in the resources spent for the word recognition task in noisy conditions. The authors concluded that the increased difficulty of the word recognition task forced the listeners to spend fewer resources on the visual task thereby increasing the reaction time for visual stimulus. Rabbit (1966) used a digit recall testing method to determine whether items that were difficult to recognize were also difficult to remember. A decrease in the recall ability for digits in the presence of noise was found and it was demonstrated that increased difficulty in the recognition of speech in noisy environments would interfere with the performance of other simultaneous activities. Apart from using word recognition scores during simultaneous task environments, other subjective measures such as Subjective Workload Assessment Technique (SWAT) and NASA-Task Load Index (NASA-TLX) could also be used to measure the workload on the listener (Horberry, Anderson, Regan, Triggs, & Brown, 2006; Lee, Vaven, Haake, & Brown, 2001; Matthews, Legg, & Charlton, 2003; Rubio, Diaz, Martin, & Puente, 2004).

Subjective Workload Assessment Technique (SWAT).

The subjective workload assessment technique (SWAT) has a multidimensional approach of measuring the perceived workload. The different dimensions of SWAT were: time load, mental effort load, and psychological stress load (Reid, Shingledecker, & Effemeier, 1981). Workload measurement using SWAT is based on the premise that participants can accurately perceive different amounts of workload for different tasks and the

perceived workload is a combination of the three dimensions of the SWAT measurement scale. Time load measures the contributions of time constraints and interruptions that occurred during the performance of the task. Mental load measures the contribution of difficulty and uncertainty of the task. Psychological stress load measures the frustration, confusion, and the uneasiness felt by the participant while performing the task.

Extensive research has been conducted on SWAT to measure its reliability. Charlton (1986) concluded that SWAT is a reliable and valid measurement of workload. Requiring to pre-train the participants before administering the test and the time taken to set-up the test are the disadvantages of SWAT (Charlton, 1996). Also, SWAT is considered to be less sensitive than the NASA-TLX measure (Nygren, 1991). Whitaker, Peters, and Garinther (1990) used SWAT to determine the performance on a route exercise when the directions were presented in different amount of noise. Subjects rated the difficulty of the task using SWAT questionnaire while hearing words from the Modified Rhyme Test. The results showed that increased intelligibility led to a decrease in the perceived difficulty of the level of the simultaneous task.

NASA - Task Load Index (NASA-TLX).

The National Aeronautical Space Administration (NASA) Task Load Index (TLX) (Hart and Staveland, 1988) is a multidimensional approach similar to SWAT to measure the perceived workload. NASA-TLX uses six

scales with 20 levels to measure the perceived workload. The six dimensions of the NASA-TLX questionnaire are: mental demand, physical demand, temporal demand, performance, effort, and frustration. The mental demand and temporal demand measures the same factor as indicated by mental load and time load measures of SWAT. Psychological stress load of SWAT measures the combinative effect of effort and frustration on the perceived workload. However, in NASA-TLX, effort and frustration have individual scales to measure their contribution to the perceived workload. NASA-TLX also measures the effect of performance (self-rating of the participants about how well they performed the task) and physical demand (measuring the physical activity required to complete the task) which were not considered by SWAT.

NASA-TLX was compared with SWAT, Modified Cooper-Harper Scale (MCH) (Wierwille & Casali, 1983), and Overall Workload (OW) (Wickens, 1992) for their validity in measuring the perceived workload. It was found that of all the four tests compared, NASA-TLX had the highest validity. Principal component analysis showed higher mean factor loadings for NASA-TLX as compared to other three scales. The authors concluded that the performance and perceived workload were strongly correlated and NASA-TLX had the highest correlation with performance compared to other three measures.

Cell phone driving and work load measurement.

The National Automotive Sampling System and the National Highway Traffic System Administration firmly believe that divided attention due to cognitive demands of the cell phone tasks is the largest contributor for the accidents caused by cell phones (Goodman, Bents, Tijerina, Wierwille, Lemer, & Benel, 1997). A study by Lamble, Kauranen, Laakso, & Summala (1999) provided evidence supporting the significance of the effects of cognitive demands imposed by cell phones contributing to accidents. They found an impairment of about 0.5s in brake reaction time in responding to a closing headway situation occurred during both non-visual cognitive task and visual dialing task. Alm and Nilsson (1995) also reported an increase in brake reaction time for drivers who concurrently drove and conversed on a cell phone.

Brown, Tickner, & Simmonds (1969) investigated the interference between driving and concurrently talking on a cell phone. They asked the participants to negotiate driving through gaps on a test track which were wider or narrower than the vehicle while concurrently performing a secondary task of checking the accuracy of sentences using a hands free telephone. The authors concluded that talking on cell phone and performing simultaneous task has little impact on the performance of more automated process (driving). However, performance on other cognitive processes that

required perception and decision making were affected while the participants switched between visual and auditory tasks.

McKnight and McKnight (1993) concluded that as the complexity and intensity of cell phone conversation increased, driving performance decreased and subjective workload increased. Parkes and Hooijmeijer (1993) also found an increase in subjective work load index measured using the NASA-task load index (TLX) when a cell phone task was introduced for drivers driving on a motorway. Alm and Nilsson (1995) examined the consequences while driving, using NASA-TLX to measure the subjective work load and Baddeley working memory span as a cognitive workload. They reported a significant increase in mental demand, effort, frustration, and time pressure during the cognitive tests.

Brookhuis, De Vries, & De Waard (1991) reported an increase in heart rate and subjective workload when participants used hands free and hand held cell phones while driving. However, no effect was found for the telephone type suggesting no difference between hand-held telephone and hands free telephone on workload. They used a simple, unspecified, analog workload scale, which may not be sensible to variations between cell phone types.

Fairclough, Ashby, Ross, & Parkes, (1991) used NASA-TLX questionnaire to measure subjective workload. They measured the workload in three conditions: talking on a telephone and driving, talking to a fellow passenger and driving, and driving only conditions. The NASA-TLX

questionnaire results revealed that there was a higher workload associated with the first two conditions compared to the driving only condition. During both conversation conditions, they noticed a reduction in speed which may present a conscious or unconscious attempt by the participant to reduce the demands of driving task which would in turn reduce the subjective work load.

Rakauskas, Gugerty, & Ward (2004) investigated the effects of easy and difficult mobile phone conversations on driving performance. They found that mobile phone conversation significantly changed driving behaviors in terms of increased accelerator pedal position variability, increased speed variability and reduced average driving speed. They also reported a higher level of workload regardless of conversation difficulty level. They concluded that drivers coped with additional stress of phone conversations by enduring higher workloads or by setting reduced performance goals.

The variety of results reported indicate the complexity of the driving task as well as the large number of dimensions that require a driver's attention. It will be argued that another dimension has been neglected until the present work. This additional challenge is based on the quality of the speech presented to drivers. Specifically, the quality of cell phone speech, computer synthesized speech, and speech in noise might be expected to shift attention from driving, but this has never been investigated.

Speech Perception and Other Tasks.

The task of driving and talking combines the issues of speech quality, speech content, and simultaneous visual-motor performance. The effect of content on speech perception has been studied extensively (Marsden-Johnson, 1990). However, the importance of the quality of the speech transmitted on speech perception and performance on a simultaneous visual-motor task has not been studied. Using a cell phone concurrently with driving shares limited attention capacity. According to the single resource theory, attention is a single limited resource. When the combined attention demands by using cell phone while driving exceeds the attention capacity, performance degradation on either task is expected. Tarawneh (1991) defined the driving task as a combination of sensory, mental, and motor processes. The sensory process involves sensing and scanning the environment. The mental process involves perception, decision making, and feedback control. The motor process involves the initiation of reaction. The driving task as defined Tarawneh (1991) fits the theory of situation awareness well. Endsley (1988) defined situation awareness as a person's perception of the elements of the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future. There is a variation is the performance degradation of using cell phone while driving due to the differences in the driver's abilities and skills, driving conditions,

and the various levels of complexity in the secondary task of using the cell phone.

While limited capacity does not imply an inability to do two or more tasks simultaneously, it means that several demanding tasks cannot be carried out simultaneously without some interference. An important factor in this connection is the degree of similarity between the tasks as has been proposed by Wickens (1984) in his 'multiple resource theory' of human performance. According to this theory, it is easier to perform simultaneous sensory and motor tasks than it is to perform tasks that require the use of same modalities. Wickens (1984) proposed a three-dimensional model for the structure of processing resources, where the dimensions are stages, modalities, and codes with associated responses. If we consider speech perception to be primarily a verbal/vocal task and visual-motor task such as driving to be a spatial/manual task, the resources required for the two are well separated on the response dimension. Accordingly, there should be very little or no interference between the auditory and visual task.

Strayer, Drews, & Johnston (2003) found that drivers were more likely to miss critical traffic lights (stop signs, traffic lights, a vehicle braking in front of the driver, etc.), slower to respond to the signals that they do detect, and more likely to be involved in rear-end collisions when they are conversing on a cell phone. The authors concluded that there was no direct relationship between gazing at passing objects and actually seeing the objects in a driving

environment. Reduced ability to see objects in driving environment was attributed to the reduction in the visual inputs available. According to the authors, while talking on a cell phone and driving, the participants' attention was directed at an "internal, cognitive context associated with the phone conversation" as compared to the external distracters in the driving environment.

Prior research have established that driving and simultaneously talking on a cell phone increases the reaction time of the driver for initiating a braking response (Alm & Nilsson, 1995; Irwin, Fitzgerald & Berg, 2000; Strayer & Drews, 2004; Strayer, Drews & Johnston, 2003). Also, the probability of missing a traffic signal doubled when the driver was conversing on a cell phone (Strayer and Johnston, 2001). McKnight and McKnight (1993) found that the response rate, which could be any threat reducing vehicle control input such as deceleration and braking, of younger people increased by lesser degree as compared to older people in casual conversations where as it was the same for both the age groups for intense conversations. Also, response rate was not influenced by gender and prior experience with cell phones. Several studies have found that working memory tasks (Alm & Nilsson, 1995), mental arithmetic tasks (McKnight & McKnight, 1993) and reasoning tasks (Brown, Tickner & Simmonds, 1969) deteriorate the performance in simulated driving tasks. Strayer, Drews, Crouch & Johnston (2005) found that talking on the cell phone while driving disrupt driving

performance by diverting attention from the factors associated with safe driving.

Two factors increase the chances of involving in an accident while driving and simultaneously using a cell phone (Matthews, Legg, & Charlton, 2003). The first factor is visual and physical competition between driving and using a cell phone. While placing or receiving calls, the drivers must momentarily remove their vision from the road and at least one hand from the steering wheel and look at the cell phone for operating it. The second factor contributing to involving in an accident is the cognitive competition between driving and operating a cell phone. A person's ability to concurrently do two or more tasks is generally limited to one task requiring conscious effort (controlled processing) and one or more tasks requiring little to no conscious effort (automated processing) (Schneider & Shiffrin, 1984).

Speech Communication and Driving

Listening to speech while performing simultaneous tasks is probably more common than listening to speech alone (Tun & Wingfield, 1994). A modern variant of this situation is listening to speech via a cell phone while simultaneously driving a vehicle. During the past 10 years, cell phones have gained substantial worldwide acceptance. Since early 1984, when the first complete cell phone systems became operational, their popularity has continued to rise. In the US alone, there were an estimated 270 million cell phone subscribers as of December, 2008. The total number was expected to

surpass three million by the end of 2009 (CTIA 2008). This increase in the number of cell phone users has increased the number of people who simultaneously talk and drive. For example, surveys indicate that 85% of cell phone owners use their phone occasionally while driving. Moreover, 27% of the people use their phones half of their total trip time of more (Goodman, Bents, Tijerina, Wierwille, Lemer, & Benel, 1997; Goodman, Tijerina, Bents, & Wierwille, 1999). Although comprehensive effects of cell phone usage on public safety are not clearly known, as much as half of motor vehicle accidents on US highway have been linked to driver inattention and other human errors (US Department of Transportation, 1998). Due to the possible increase in the risks associated with the use of cell phone usage while driving, several legislative efforts have been made to restrict the use of cell phones while driving.

699 individuals who had cell phones and involved in motor vehicle accidents resulting in high property damage and no personal injury were studied to arrive at the relationship between cell phone use and motor accidents (Redelmeirer and Tibshirani, 1997). It was found that the release of cognitive resources used during cell phone conversation was not instantaneous. 24% of the drivers involved in the accidents had used their cell phones within 10 minutes before the accident. The authors claimed that the likelihood of the driver to get involved in an accident increased four times while using the cell phone and driving simultaneously. Also, the relative risk

of talking on the cell phone and driving was similar to that of driving with a blood alcohol level above the legal limit. The authors found no reliable advantages of using a hands-free equipment over the traditional hand-held equipment and the chances of getting involved in an accident were the same for both the sexes and all age groups. The authors concluded that the interference associated with cell phones was due to lack of attention while driving and talking simultaneously on cell phone than to other factors such as holding the cell phone, making a call, etc.

Despite this interpretation, other research has reported conflicting findings. For example, two studies have shown that manual manipulation of cell phones such as dialing the phone and answering the phone had a negative impact on driving (Briem & Hedman, 1999; Strayer & Johnston, 2001). Briem and Hedman discovered that simply conversing over a hands free telephone while driving does not impair performance. Clearly there were differences between the studies that show that the effects of conversation on driving are not yet fully understood. Perhaps, a difficult conversation affects driving and any prolonged manipulation of the telephone may produce a decrease in performance. Perhaps the findings of these experiments could be reconciled if they were evaluated in light of the difficulty of the nature of conversation and the mechanical manipulation of the telephone and radio buttons.

A third possibility that might increase the chances of involving in an accident while driving and simultaneously using a cell phone has never been investigated. It is possible that low-quality speech interferes with simultaneous visual-motor task. A variety of related studies make this a reasonable possibility. It is well established that vocoded speech is less intelligible than natural speech (Altom, Macchi, & Wallace, 1990; Greenspan, Bennett, & Syrdal, 1994; Silverman, Kalyanswamy, Silverman, Basson, & Yashchin, 1993; Spiegel, Van Santen, 1993; Venkatagiri, 2003; Winters, & Pisoni, 2004). Strayer & Johnston (2001) found that drivers had difficulty in perceiving both natural and vocoded speech while driving, the difficulty being the highest for vocoded speech. Jamieson et al (1996) found that listeners with normal hearing have significantly more difficulty understanding speech when it has been processed by certain coding schemes. Choi (2005) found that effect of vocoded speech on a simultaneous visual-motor task is bidirectional. Listening to vocoded speech decreased the performance on a visual-motor task as much as affecting the perception of speech itself.

Chapter III

Rationale

Statement of the Problem

Traditional intelligibility measures were originally developed to determine the effectiveness of communication systems. Soon after, they were adapted to evaluate the integrity of the auditory system and estimate the communication efficiency of different kinds of speech at different environments (Lavandier and Culling, 2010; Longworth-Reed, Brandewie, and Zahorik, 2009; Yorkston and Beukelman, 1991). Also, intelligibility measures are commonly used to validate the appropriateness of different signal processing techniques applied to speech (Carrell and Opie, 1992; Edwards, 2000; Picheny, 1985; Villchur, 1973). In general, intelligibility measures are often compared within the same kinds of speech (intelligibility at different SNRs) or across different kinds of speech (natural speech versus cell phone speech, speech from two different synthesizers) to measure their advantages and drawbacks (Francis, Nusbaum, and Fenn, 2007; Mirenda and Beukelman, 1990; Logan, Greene, and Pisoni, 1989). Although traditional intelligibility measures have been well accepted by researchers, they have also been criticized for their low sensitivity in detecting the effects of quality of the stimuli on performance (Brungart, 2001; Jerger and Jerger, 1983; Pratt, 1981). For example, two different kinds of speech at different SNRs may end up having same intelligibility as reported by the individuals.

Furthermore, intelligibility measures have also been shown to have low validity in predicting communication ability in real life even after the reliability of the test is well controlled (Walden, 1984).

Despite some drawbacks, traditional intelligibility measures have been used to evaluate different kinds of speech at multitude of environments. In fact, most of the synthetic speech studies that compared synthetic speech produced by different synthesizers used intelligibility measure as their testing criteria (Logan, Greene, & Pisoni 1989; Raghavendra & Allen, 1993; Schwab, Nusbaum, & Pisoni, 1985; Stevens, Lees, Vonwiller, & Burnham, 2003). However, the percent correct measure arrived from an intelligibility test is not a simple indicator of the intelligibility inherent with the attributes of the target stimuli. Rather, the resulting performance in a given intelligibility test is a combined result of stimulation at the sensory organ caused by the acoustic signal with a certain physical characteristics and judgment based on sensory input that is reconstructed of modified to find a best match and compare with the pattern stored in the long-term memory (Pisoni, 1982). The resulting performance through a speech intelligibility test is combination of sensory processes and other cognitive processes pertaining to the incoming information. For instance, when the input stimulus is free from noise and other distortions that might happen due to delay and reverberation and has no ambiguity underlying in the signal, the perception of the stimuli may be similar across all listeners. When there is a high

amount of ambiguity involved at the input stage, individual listeners' internal resources can play a significant role and influence the outcome of the experiment. Listeners use more cognitive processing and effort to make best sense of the stimuli in noise presented to them (Pisoni, 1982). Information processing models of speech perception predict that listeners use prior semantic, syntactic, and lexical knowledge to fill the missing information in the original signal. These additional processes require resources to temporarily store the information extracted until all the processes are finished for the complete perception of the presented signal. Since we cannot directly measure the processing demands and special strategies employed by a listener to perceive an ambiguous message, we must indirectly predict the processing demands by measuring how much attentional effort is exerted on these secondary processes to achieve a certain level of performance. With traditional intelligibility measures, these differences in processing demands have not been measured. This is because with traditional intelligibility measures, increased processing demands may not affect the overall performance on the intelligibility test. By doing multiple tasks simultaneously, the listener may be mentally fatigued after a while listening to speech in a noisy background. Moreover, attention plays an important role in speech perception. Most of the speech perception and word recognition theories do not have models that incorporate the effect of fatigue and other simultaneous tasks. Also, measuring the increased processing demands

caused by listening to different kinds of speech in a noisy environment may be as important as to predicting how people listen to different kinds of speech in a noisy environment while performing multiple simultaneous tasks.

Intelligibility alone is not sufficient to evaluate how different kinds of speech are perceived by listeners while doing other tasks simultaneously. Other criteria should be considered and developed to completely understand the phenomenon of speech perception while performing simultaneous tasks. The simultaneous tasks should be incorporated in such a way that they impose different levels of cognitive demands on the listener. Hence simultaneous task performance measured by tracking performance and cognitive load measured by visual word identification task were added to the experiments. The results from these additional measures may help us to determine the different processing strategies used by the listener to perceive different kinds of speech in various attentional environments. In the present study, a non-speech intelligibility based measure focusing on increased processing demands as indication of increased listening effort was explored to evaluate the effect of simultaneous tasks in perceiving different kinds of speech.

Overall Purpose

Due to limitations in the measure of intelligibility such as failure to reveal the effort required by a listener to reach a particular level and the inability to provide a unidimensional measure for different speech sources

and listening environments; an additional method was sought. To evaluate the effects of performing simultaneous tasks on the perception of different kinds of speech, performance on the simultaneous visual-motor task and performance on a visual word identification task were evaluated.

Three experiments were conducted to investigate the effects of performing simultaneous visual-motor tasks on the perception of different kinds of speech. The first two experiments used a dual-task paradigm with an intelligibility task and an adaptive visual-motor task. The third experiment used a multi-task paradigm comprised of an intelligibility task, an adaptive visual-motor task, and a visual word identification task. The visual word identification task was used to investigate how a theory of attention might explain the interaction between speech perception and visual-motor task performance.

Research Questions

The following research questions were asked to investigate the effects of performing simultaneous tasks on the perception of different kinds of speech in the series of three experiments.

Experiment I – Comparison of differences of speech sources

1. Does performance on the visual-motor task differentially affect the perception of natural speech, synthetic speech and cell phone speech?

Experiment II – The effects of noise on different speech sources

1. How does the perception of different speech sources change as a function of background noise?

    a. Does the signal-to-noise ratio level affect intelligibility?
    b. Does the signal to noise ratio level affect visual-motor performance?
    c. Do intelligibility and visual-motor tasks interact when performed simultaneously?

2. Does semantic context affect performance on simultaneous motor and word repetition tasks?

3. Do different speech sources impose different cognitive load?

Experiment III – What explains the differences in the perception of sound sources

1. Do different speech source qualities use different attention mechanisms?

    a. Are speech sources (e.g., natural versus cell phone speech) processed differently during simultaneous consistent mapping versus varied mapping tasks?

2. Are speech and visual-motor tasks affected differentially by consistent mapping versus varied mapping tasks?

3. How is the simultaneous visual-motor task affected by adding additional visual word recognition and memory tasks?

4. Does the performance on visual word identification task differ while listening to cell phone and natural speech during CM and VM tasks?

CHAPTER IV

Experiment 1

Experiment 1 was designed to contrast the effect of a simultaneous visual-motor task on the intelligibility of three types of speech: synthetic, cell phone, and natural. Participants were simultaneously presented with a series of spoken sentences via circumaural headphones and moving target on a video display. They were instructed to repeat the last word of the sentence and to keep a cursor on top of the moving target dot on the computer monitor. The participants were told to perform each of the tasks as rapidly and accurately as possible. In addition to the three types of speech sources (natural, cell phone, and synthetic), the acoustic stimuli were presented in two levels of background noise. Pilot experiments were conducted to arrive at reasonable signal-to-noise ratios of the stimuli and adaptive pursuit rotor parameters.

It was hoped that the pattern of results from this array of stimuli would reveal differences and similarities in perceiving speech produced by these different sources. For example, listening to natural speech in loud noise may have the same effect on target-following as listening to synthetic speech in a quieter background. On the other hand, it might be found that synthetic speech always leads to poor target following, no matter what the noise background. These different results would provide two very different models of speech perception. Moreover, the different models would suggest

alternative approaches to general speech processing issues in human-computer interaction, AAC, hearing aids, and cochlear implants.

Participants

Twenty-four undergraduate students in the Special Education and Communication Disorders Department at University of Nebraska - Lincoln volunteered for this study. They ranged in age from 19 to 26 with an average age of 21. The participants were given a hearing screening to verify audibility no poorer than -20 dB @500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. The participants were randomly assigned to a counterbalancing condition which determined the order of stimulus presentation.

Stimuli and Apparatus

Adaptive Pursuit Rotor.

Variations of the pursuit rotor task (Snoddy, 1926) have been used widely since its development. In the present work, an adaptive variant of the traditional pursuit rotor task called 'adaptive pursuit rotor' (Srinivasan and Carrell, 2006) was used to measure participants performance while performing simultaneous visual motor task. Adaptive pursuit rotor allows people with wide variety of skills and simultaneous task experiences to be tested reliably.

Participants performed an adaptive pursuit tracking task as the simultaneous task with speech perception. In the adaptive pursuit rotor task,

the participants used a stylus to move a cursor on a computer display to keep it aligned as closely as possible to a moving target. Similar pursuit rotor tasks have been used extensively to study the underlying mechanisms of motor learning (Siegel, 1990). In the present experiment, however, it was used as a distractor from the primary speech repetition task. The Adaptive Pursuit rotor (APR) task involves following a moving target on a computer screen using a stylus to keep the cursor aligned as closely as possible to the moving target. Figure 2 is a screen capture from the adaptive pursuit rotor task.

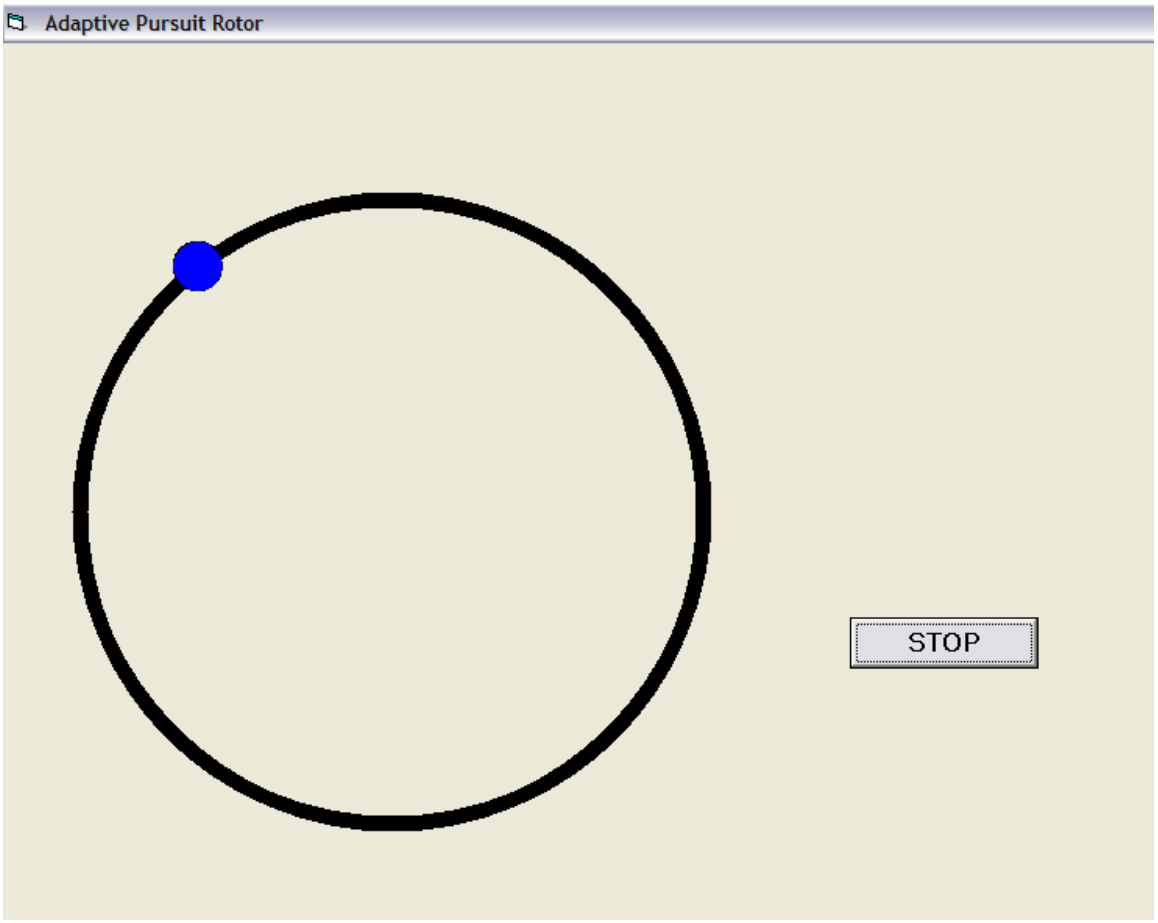

Figure 2. Adaptive Pursuit Rotor screen

The cursor position was controlled by a stylus and a track pad on a desk. The target position was a point in a circle of fixed radius. The target motion was smooth and continuous. If the participant was on target for at least 70% of the time, the target increased its speed. If the participant was not on target for at least 70% of the time, the target decreased its speed. For example, if the speed of rotation is 5 rotations per minute and the participant was on target for 65% of the time, the speed of rotation was reduced to 4.5 rotations per minute. On the other hand, if the participant was on target for 80% of the time, the speed of rotation was increased to 5.5 rotations per minute. The target position was updated adaptively using a moving average filter so that the time on target resulted in an average performance of 70 percent correct throughout the experiment.

Traditional pursuit rotor experiments had problems with participants learning the task very well (Noguchi et al., 2005). In order to keep cognitive workload high throughout the experiment and to have the participants start at a similar level of expertise irrespective of previous experience, the simultaneous pursuit rotor task was made adaptive.

Stimuli.

All sentences were taken from the Speech Perception in Noise (SPIN) set sentences (Kalikow, Stevens, & Elliot, 1977). The stimuli were recorded in a background of multitalker babble (Bilger, R. C., et al, 1984). One set of sentences was naturally produced, another set was recorded via a cell phone

link and a final set was synthetic speech from AT&T Natural Voices (TTS Publications. (n.d). Retrieved February 22, 2006, from http://www.research.att.com/projects/tts/index.html).

SPIN sentences consist of two types of sentences; sentences with predictable final words, and sentences with unpredictable final words. In predictable last-word sentences, the last word of the sentence is selected from a closed set of words. For example, in the sentence "The old train was powered by steam", the last word can be from a closed set of words which describes how the train was powered such as steam, diesel, electricity, etc. Any other words not from this closed set would not be semantically correct. In unpredictable last-word sentences, the last word of the sentence is from an open set of words. For example, in the sentence "Jack was thinking about the car", the last word can be nearly any word from the lexicon. Any word from the lexicon will make the sentence semantically and syntactically correct. SPIN sentences were selected because they allow the contributions of both semantic and phonetic information to be evaluated.

The naturally-produced speech was spoken by a female speaker with a General American English dialect in a sound treated acoustic chamber with an A-weighted background noise level of 18 dB SPL. The sentences were spoken into a Crown CM312 microphone situated at a distance of 2 cm from the side of the mouth. The recorded sentences were then digitized with a sampling rate of 44100 samples per second and quantized at 16 bits per

sample. After digitization, the sentences were downsampled to 11025 Hz and low-pass filtered with a cut-off frequency of 4500 Hz. They were then segmented into individual files for later organization

The cell phone sentences were derived from the same natural tokens of SPIN sentences. The natural speech was downsampled to 11025 Hz and low-pass filtered with a cut-off frequency of 4500 Hz. This signal was played into a Motorola V 60 cell phone as follows. The cell phone dialed the Speech Perception Laboratory wireline telephone through the Public switched Telephone Network (PSTN). The laboratory phone was connected to a TDT A/D converter using a 600 ohm – 600 ohm isolation transformer and a 4500 Hz low-pass filter. The sampling rate of the A/D converter was 44100 Hz. The recorded sentences were then downsampled to 11025 Hz, low-pass filtered with a cut-off frequency of 4500 Hz and was segmented into individual sentences for later organization

The stimuli were further divided into predictable and unpredictable sentences and the order in which these stimuli were presented was counterbalanced using a Latin square design. The exact arrangement is shown in figure 4.

Apparatus.

Two experimental-control programs, IDConnect (Srinivasan and Carrell, 2005) and APRConnect (Srinivasan and Carrell, 2006), were used to control the experimental procedure. IDConnect was used to present the

sentences to the participants over the headphones (Sennheiser HD 520). APRConnect was used to measure the work load increment in dual task condition and to measure the interference caused due to listening effort. Both these programs were synchronized to run simultaneously via a network connection. Figure 3 shows the connections necessary for synchronizing between the two programs and controlling the experiment.

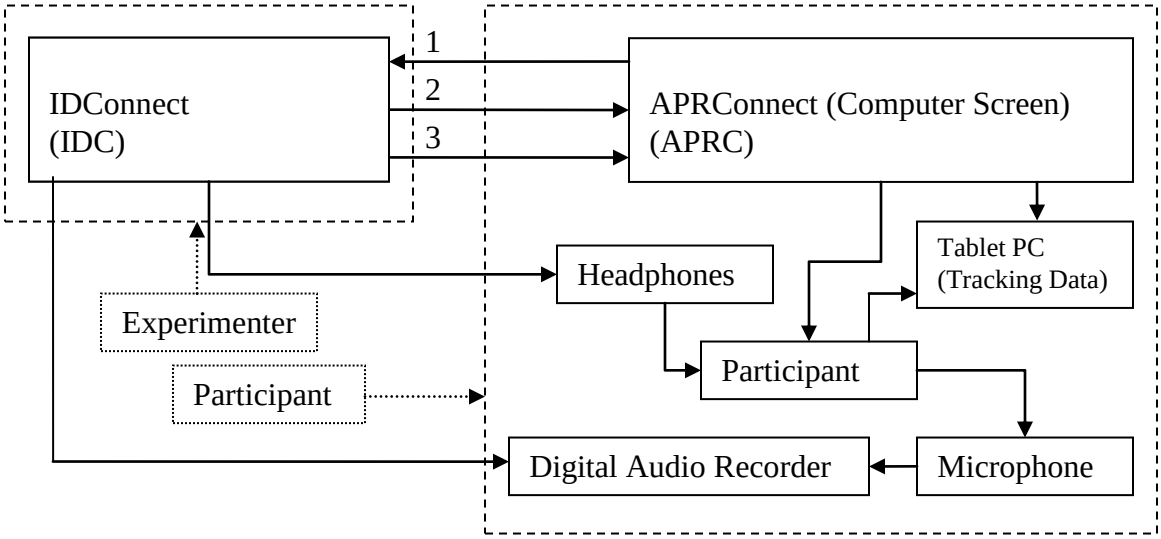

Figure 3. Synchronization between IDC and APRC

When the participant is ready, APRC sends a signal to IDC (Signal 1). When IDC receives signal 1, it starts playing the stimulus. IDC sends signal to APRC after the end of individual sentences (Signal 2). This is used to evaluate the participant's performance for individual sentences. At the end of every sentence, the participants repeated the last word of the sentence which was recorded using a digital audio recorder. When all the sentences have been played, IDC sends a signal to APRC (Signal 3) so that it could stop

collecting and save the tracking data for future analyses. Stimuli presented to the participants were recorded on one channel and participant's verbal responses were recorded on the other channel of the recorder. Figure 4 shows the sequence of events in each trial.

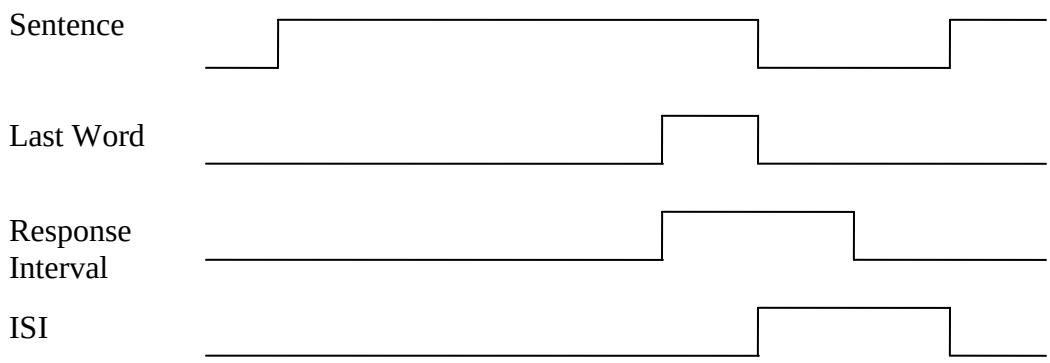

Figure 4. State diagram of SPIN sentence trial

Procedure

When participants arrived for the experiment, they completed a questionnaire assessing their hearing status, cell phone usage and mouse usage, demographic information, and their driving habits with cell phone. Participants were then familiarized with the adaptive pursuit rotor task. During this demonstration phase, the participants heard the experimenter explaining the experimental procedure and watched the experimenter doing the adaptive pursuit rotor task.

The study consisted of two phases. The first phase, familiarization, lasted for about 3 minutes and was used to acquaint the participants with the

adaptive pursuit rotor task. During this phase of the experiment, the participants practiced the adaptive pursuit tracking task and discussed any questions regarding the experimental procedure with the experimenter. No data was collected during this session.

The second phase was the dual-task phase of the study. The SPIN sentences were presented to the participants via circumaural headphones at approximately 68 dB SPL. The participants were required to repeat the last word of the sentence which they just heard. While doing the listening task, the participants had to concurrently perform the adaptive pursuit rotor task. Instructions given to the participant by the experimenter were given in Appendix 1.

Dependent Measures

During the experiment, APRConnect collected the tracking data which consisted of performance of participants on the tracking task. Also, a digital recorder was used to record the sentences and the participant's vocal responses in two different channels. From this data, the following three performance variables were extracted:

Average Speed: Average speed was the average speed in rotations per minute at which the participant did the tracking task.

Reaction Time: Reaction time was measured as the time elapsed between the end of the sentence and to the moment when the participant starts answering. This was measured through visual inspection of the

waveform to reduce the variability introduced by automatic detection of speech onsets.

Word Accuracy: Percent correct represented the proportion of sentence final words correctly identified by the participant. Note that identifying the last word to be a plural when it was actually singular and vice versa were considered to be incorrect responses.

Design

The experiment was a 2 X 2 X 2 mixed factor (source by context by S/N) design. Source and context were within-subject factors and signal-to-noise ratio was a between-subject factor. Figure 4 illustrates the levels in the design of the experiment. The order in which the participants heard the different kinds of speech stimuli were counterbalanced using a modified Latin-square design to remove order effects.

Participants also were presented with natural speech. Those results were considered separately and will be analyzed separately because they were not presented at signal-to-noise ratios that the cell and synthetic stimuli were. Data was desired on natural speech but signal-to-noise ratios needed to be adjusted to avoid ceiling effects. Figure 5 shows the experimental design used for the Experiment 1.

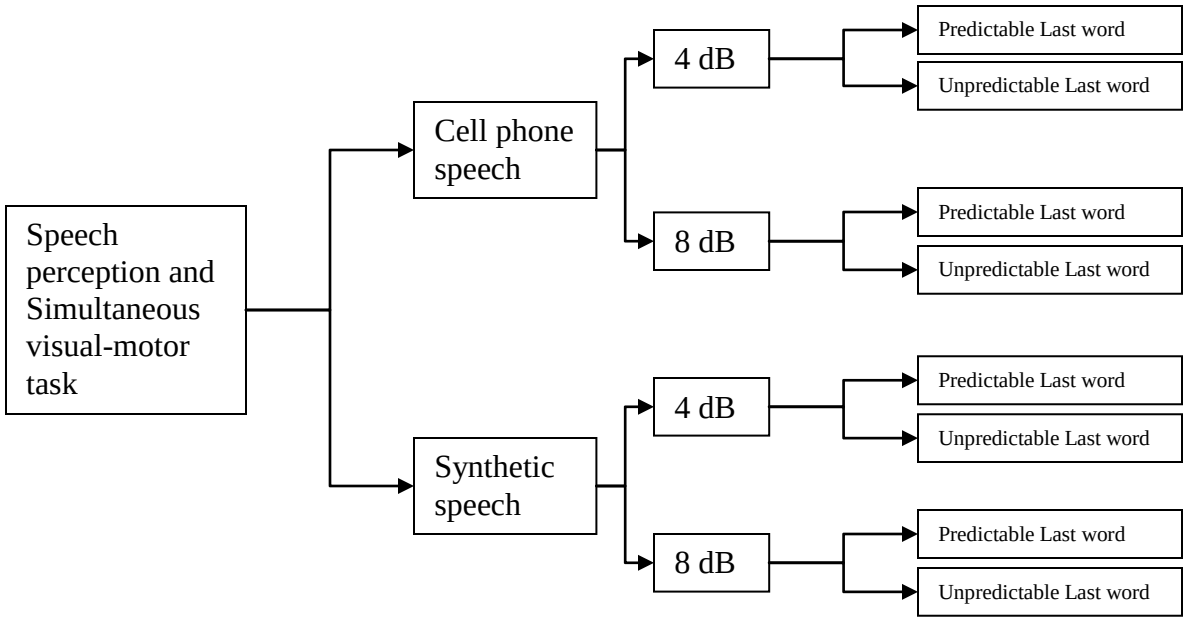

Figure 5. Experimental Design – Experiment 1

The following research questions were answered with the help of data collected with experiment 1.

1. Does performance on the visual-motor task differentially affect the perception of natural speech, synthetic speech and cell phone speech?

2. How does the semantic context of the stimulus presented (predictable sentences versus unpredictable sentences) affect synthetic speech and cell phone speech in an auditory word identification task?

Results

Average Speed.

Average speed was measured as the average speed in rotations per minute at which the participant performed the adaptive tracking task. Table 3 presents a summary of means and standard deviations for the average speed in all experimental conditions. A 2 X 2 X 2 analysis of variance, with repeated measures on the source and the context factors was performed on the data. Average speed of rotation for cell phone speech was higher than that of synthetic speech ($F(1, 23) = 82.4$, $p < 0.001$). Also, a significant main effect for level ($F(1, 23) = 23.9$, $p < 0.001$) demonstrated that increased signal-to-noise ratios increases average speed. This might be due to the fact that at higher signal-to-noise ratio levels, speech presented was more intelligible and the participants spent equal amount of resources for both auditory and visual task. At low levels of signal-to-noise ratio, speech presented was less intelligible and participants might have spent more resources on adaptive tracking task. This was evident from the word accuracy of cell phone speech at 4 dB and 8 dB signal-to-noise ratios.

The unpredictable final word sentences led to higher average speeds compared to the predictable final word sentences. This was evident from the significant source-by-context interaction ($F(1, 23) = 47.4$, $p < 0.001$). This difference was high for the cell phone speech compared to the synthetic speech. Also, the 4 dB sentences had a higher average speed compared to the

8 dB sentences. This difference in average speed was high for cell phone speech compared to synthetic speech. This was shown by the significant source by level interaction ($F(1, 23) = 28.2$, $p < 0.001$).

Table 1 shows the average speed of rotation for cell phone and synthetic speech at 4 dB and 8 dB signal-to-noise ratios. The interactions are visualized in figure 6. The error bars denote 95% confidence intervals.

Table 1. Means for the average rotation speed (rotations per minute) in visual-motor task for cell phone speech and natural speech at 4 dB and 8 dB signal-to-noise ratios for predictable and unpredictable sentences

	Cell Phone Speech		Synthetic Speech	
	4 dB	8 dB	4 dB	8 dB
Predictable	9.63	8	9.07	9.96
Unpredictable	11.08	10.1	9.92	9.73

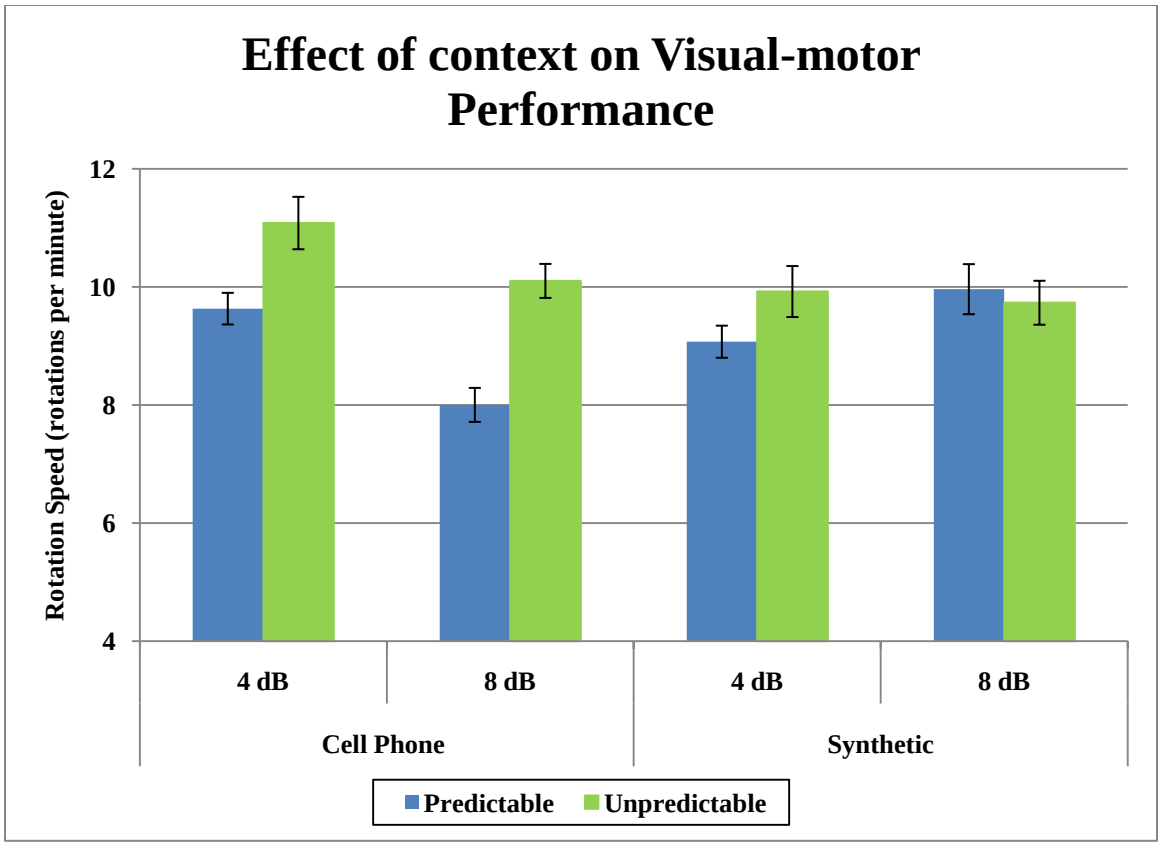

Figure 6. Effect of context on visual-motor performance for cell phone and synthetic speech

Reaction time.

Reaction time was measured as the time elapsed between the end of the sentence and to the beginning of the spoken response. Only correct responses are used in the table, figure, and resulting statistical analysis. Table 2 presents a summary of means and standard deviations for the reaction time in the dell phone and synthetic speech conditions. A 2 X 2 X 2 analysis of variance, with repeated measures on the source and context factors was performed on the data. Natural speech at 0 dB signal-to-noise

ratio and -4 dB signal-to-noise ratio were not included in the analysis because there were no comparable levels in the synthetic speech and cell phone speech conditions. Reaction time for cell phone speech was significantly lower than that of synthetic speech ($F(1, 23) = 6.2$, $p < 0.02$). A significant main effect was also found for context ($F(1, 23) = 66.2$, $p < 0.001$), indicating that the predictable final word sentences were answered faster than the unpredictable final word sentences. Also, a significant main effect was found for level ($F(1, 23) = 24.2$, $p < 0.001$), demonstrating that sentences at high signal-to-noise ratios were responded to more rapidly than the sentences at low signal-to-noise ratio.

As may be seen from the data listed in Table 1, the predictability of the final word of the sentence helped cell phone speech more than synthetic speech. This was evident from the significant source by context interaction ($F(1, 23) = 16.6$, $p < 0.001$). Also, increased signal-to-noise ratio helped cell phone speech more than synthetic speech. This was evident from the significant source by level interaction ($F(1, 23) = 11.5$, $p < 0.002$). As may be seen from the data listed in Table 1, increased signal-to-noise ratios led to decreased reaction times and this change in reaction time was almost the same for both predictable and unpredictable sentences ($F[1,23] = 6.0$, $p < 0.022$).

Table 2 shows the mean reaction time for auditory word repetition task for cell phone and synthetic speech at 4 dB and 8 dB signal-to-noise ratios.

The interactions are better visualized in figure 7. The error bars denote 95% confidence intervals.

Table 2. Mean reaction time (ms) for auditory word repetition task for cell phone and synthetic speech at 4 dB and 8 dB signal-to-noise ratios for predictable and unpredictable sentences.

	Cell Phone Speech		Synthetic Speech	
	4 dB	8 dB	4 dB	8 dB
Predictable	408	337	499	477
Unpredictable	698	467	594	566

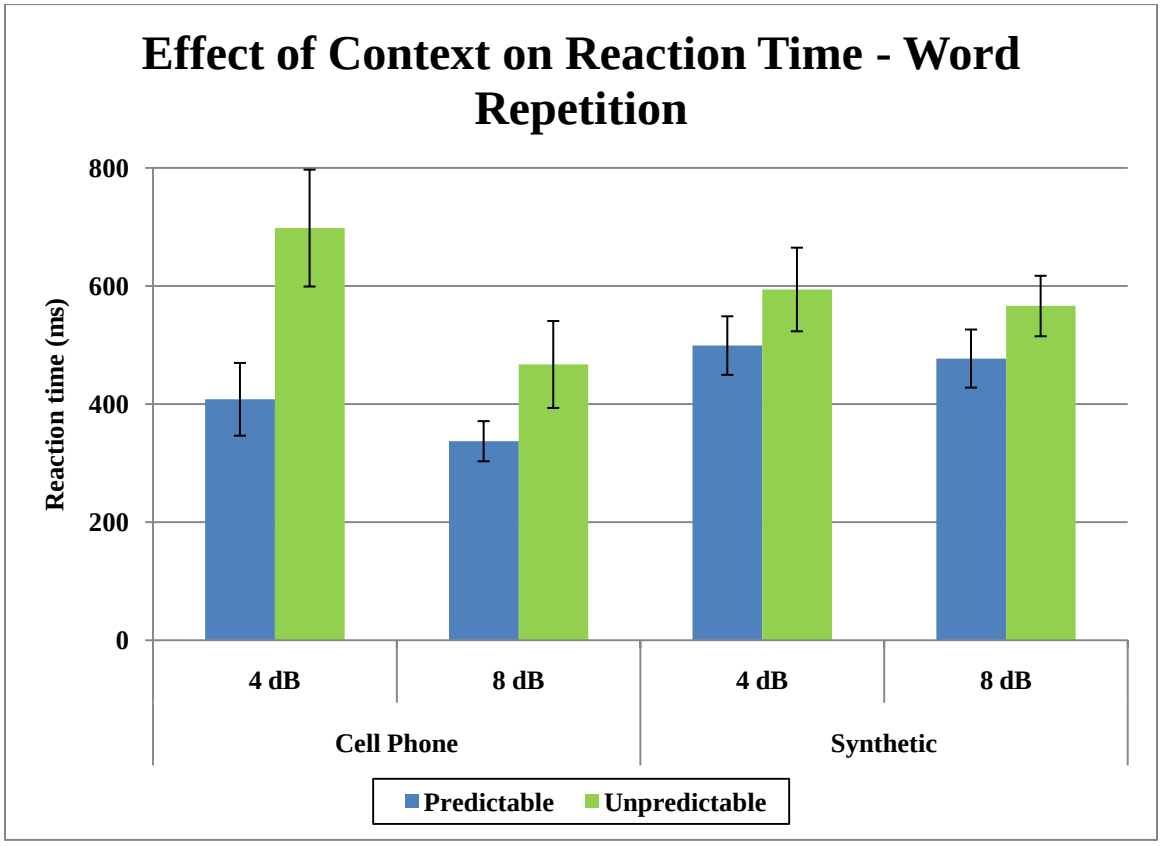

Figure 7. Effect of context on reaction time for word repetition for cell phone and synthetic speech

Word Accuracy.

Word accuracy was measured as the proportion of the final words of the sentences correctly identified by the participant. Table 2 presents a summary of means and standard deviations for the word accuracy in cell phone and synthetic speech conditions. A 2 X 2 X 2 analysis of variance, with repeated measures on the source and the context factors was performed on the data. Word accuracy for synthetic speech was higher than for cell phone speech ($F(1, 23) = 404.5$, $p < 0.001$). Higher signal-to-noise ratios increased intelligibility which was evident from the significant interaction ($F(1, 23) = 124.8$, $p < 0.001$). Also, predictable last-word sentences were more intelligible than unpredictable last-word sentences ($F(1, 23) = 552.0$, $p < 0.001$).

As may be seen from the data listed in Table 2, context improved word accuracy for cell phone speech more than synthetic speech. This was evident from the significant source-by-context interaction ($F(1, 23) = 49.0$, $p < 0.001$). Also, increased signal-to-noise ratios helped cell phone speech more than it helped synthetic speech in identifying the last word of the sentence. This was obvious from the significant source-by-level interaction ($F(1, 23) = 37.0$, $p < 0.001$). Also, the interaction of context by level was significant ($F(1, 23) = 12.4$, $p < 0.002$]. As may be seen from the data listed in Table 3, increased signal-to-noise ratios led to increase in word accuracy and this change in word accuracy was more pronounced on the unpredictable last word sentences as compared to the predictable last word sentences. The

interactions are better visualized in figure 8. The error bars denote 95% confidence intervals.

Table 3. Means for word accuracy (%) for cell phone and synthetic speech at 4 dB and 8 dB signal-to-noise ratios for predictable and unpredictable sentences

	Cell Phone Speech		Synthetic Speech	
	4 dB	8 dB	4 dB	8 dB
Predictable	71.67	90.42	90.62	94.37
Unpredictable	23.95	51.04	63.75	70.04

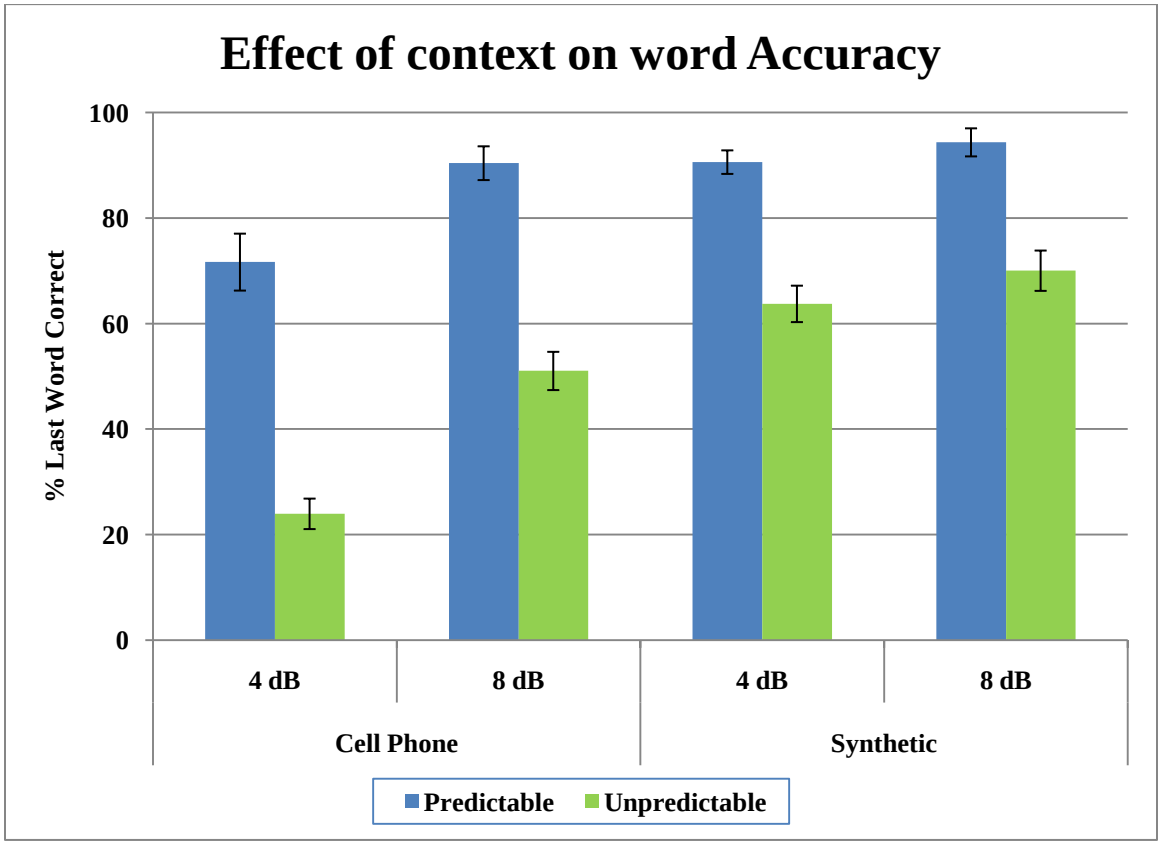

Figure 8. Effect of context on word accuracy for cell phone and synthetic speech

Though experiment 1 quantified listeners' performance on speech intelligibility and reaction time while performing a simultaneous task, these results were of limited importance. Because only two levels of signal-to-noise ratios were presented, it was impossible to derive a complete relationship between noise and performance on the pursuit rotor task. That is, the relation between a gradual change in signal-to-noise ratio in a speech stimuli and the performance in a simultaneous visual-motor task could not be quantified from the results of experiment 1. Also, experiment 1 did not provide information about strategies used by the participants. In order to refine the results from Experiment 1, a second experiment was designed. Cell phone speech and natural speech were presented at many different signal-to-noise ratios so that the strategies used by the participants at different levels could be investigated and more complete relationships between signal-to-noise and performance could be demonstrated.

CHAPTER V

Experiment 2

Experiment 2 was designed, in part, to allow natural speech to be systematically compared with cell phone speech and assist in developing a more comprehensive explanation of the data from experiment 1. Specifically, experiment 2 examined the effects of presenting sentences across a range of signal-to-noise ratios using cell phone speech and natural speech on intelligibility and pursuit rotor performance. The major difference between Experiments 1 and 2 was that the natural speech and the cell phone speech were presented across a range of different signal-to-noise ratios in Experiment 2 as compared to Experiment 1 where just two levels were used.

The data were used to investigate the effect of this wider range of signal-to-noise ratios on the speed of rotation for cell phone and natural speech sentences. As the participants performed a simultaneous task, either the natural or cell phone speech was presented via headphones. Listeners were instructed to repeat the last word of the sentence while performing a simultaneous task. After the end of every trial, the participants rated the perceived cognitive load for the task just completed using a computerized version of the NASA-TLX questionnaire (Hart & Staveland, 1988). It was expected that as the signal-to-noise ratio of the signals decreased, speech intelligibility and tracking performance would decrease. Also, it was expected that as the difficulty of the cell phone speech presented increased,

performance on the simultaneous task would decrease. It was posited that there exists an optimal level of signal-to-noise ratio beyond which the listener might nearly abandons his or her auditory task and begins to allocate more resources to the simultaneous task. This would improve the overall performance on the simultaneous task.

The following research questions were answered based on the data collected with experiment 2.

1. What effect does changing the signal-to-noise ratio have on the performance in a simultaneous motor task?

2. What effect does changing the signal-to-noise ratio have on the performance in a word repetition task?

3. Does semantic context of the stimuli presented improve the performance on simultaneous motor and word repetition tasks?

4. Does different kinds of speech impose different cognitive load on the listener?

5. What effect does semantic context of the presented speech stimuli has on the performance in a word repetition task for different signal-to-noise ratios?

Participants

Thirty six undergraduate students of Special Education and Communication Disorders Department of University of Nebraska, Lincoln volunteered for this study. They ranged in age from 19 to 28 with an average

age of 21.41. The participants were given a hearing screening to verify the audibility of the tones at 500Hz, 1000 Hz, 2000 Hz and 4000 Hz. The participants were randomly assigned to a counterbalancing order in which they hear the stimuli.

Stimuli

Cell phone speech at 0 dB, 2 dB, 4 dB, 6 dB, 8 dB, 10 dB signal-to-noise ratios and natural speech at 2 dB, 0 dB, -2 dB, -4 dB, -6 dB, -8 dB signal-to-noise ratios were used for this experiment. Different levels for cell phone speech and natural speech were chosen based on the pilot data to avoid ceiling and floor effects.

The SPIN sentences were presented to the participants via circumaural headphones at approximately 68 dB SPL. The participants were required to repeat the last word of the sentence which they just heard. While doing the listening task, the participants concurrently performed the adaptive pursuit rotor task.

Procedure

When participants arrived for the experiment, they completed a questionnaire assessing their hearing status, cell phone usage and mouse usage, demographic information and their driving habits with cell phone. Participants were then familiarized with the adaptive pursuit rotor task. During this demonstration phase, the participants heard the experimenter

explaining the experimental procedure and watched the experimenter doing the adaptive pursuit rotor task.

The participants were randomly assigned to one of the two groups. The stimuli in the two groups were arranged in such a way that the participants heard one set of easy stimuli, one set of medium difficulty stimuli and one set of difficult stimuli. Participants in Group 1 were presented with cell phone speech at 0 dB, 4 dB, 8 dB signal-to-noise ratios and natural speech at 0 dB, -4 dB and -8 dB signal-to-noise ratios. Participants in Group 2 were presented with cell phone speech at 2 dB, 6 dB, 10 dB signal-to-noise ratios and natural speech at 2 dB, -2 dB and -6 dB signal-to-noise ratios. Participants heard different signal-to-noise ratios of speech in counterbalanced order to remove any order effects. Also, by presenting different levels of the same stimuli, the relationship between the signal-to-noise ratio and the speed of rotation could be better described. The participants were grouped into two groups so fatigue and boredom would be minimized. The presentation scheme is shown in the following table.

Table 4. Presentation Scheme – Experiment 2

Group →	GROUP 1		GROUP 2	
Difficulty ↓	Cell Phone Speech	Natural Speech	Cell Phone Speech	Natural Speech
Easy	8 dB	0 dB	10 dB	2 dB
Medium	4 dB	-4 dB	6 dB	-2 dB
Difficult	0 dB	-8 dB	2 dB	-6 dB

The study consisted of two phases. The first phase, familiarization, lasted for about 3 minutes and was used to acquaint the participants with the adaptive pursuit rotor task. During this phase of the experiment, the participants practiced the adaptive pursuit tracking task and discussed any concerns regarding the experimental procedure with the experimenter. No data was collected during this session.

The second phase was the dual-task phase of the study. During this phase of the experiment, the auditory stimuli were presented to the participants via circumaural headphones and were required to repeat the last word of the sentence which they just heard. While doing the listening task, the participants concurrently performed the adaptive pursuit rotor task

Dependent Measures

During the experiment, APRConnect collected the data which consisted of performance of participants on the tracking task. Also, a digital recorder was used to record the sentences and the participant's vocal responses in two different channels. From this data, the following three performance variables were extracted:

Reaction Time: Reaction time was measured acoustically as the time elapsed between the end of the SPIN sentence to the initiation of the participant's answer.

Word Accuracy: Word accuracy was the proportion of the final words of the sentences correctly identified by the participant. Identifying the last word to be a plural when it was actually singular and vice versa was considered to be an incorrect response.

Average Speed: Average speed was the average speed in rotations per minute at which the participant performed the tracking task.

Perceived Cognitive Load: A computerized version of the NASA-TLX questionnaire used to measure the cognitive load perceived by the participant. The computer program calculates the perceived cognitive load based on the responses of the participants' for the questionnaire.

Results

A 3-way within groups ANOVA was used to examine the main effects and interactions of speech (cell phone and natural), level (SN1, SN2, SN3,

SN4, SN5, and SN6 – 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs for cell phone speech and -8 dB, -6 dB, -4 dB, -2 dB, 0 dB, and 2 dB SNRs for natural speech), and context (predictable and unpredictable) as they relate to the average rotation speed in adaptive pursuit rotor task, word accuracy in the auditory word repetition task, and NASA-TLX. Since all the three way interactions, speech by level by context, were not significant, the analysis was divided into cell phone speech and natural speech for better understanding of the results.

Comprehensive Results.

The first research question investigated the relationship between the signal-to-noise ratios of the transmitted speech stimuli and the performance in the visual-motor task. To answer this question, the average rotation speed on adaptive pursuit rotor task at the different levels of presented stimuli was compared.

Rotation Speed - Cell Phone Speech

A 2-way within groups ANOVA was used to examine the main effects and interactions of level (0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs) as they relate to the average rotation speed in the adaptive pursuit rotor task when cell phone speech was presented to the participants. There was a main effect of level ($F\ (5, 85) = 48.3$, $p < 0.001$, $M_{se} = 0.561$). The pattern of mean differences was that the participants had a very high average speed of

rotation at very low SNRs. As the SNR increased, the average speed of rotation started to decrease. However, after a certain SNR, the average speed of rotation started to increase (LSD_{mmd} = 0.3). Table 5 shows the average speed of rotation during at levels 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs. Figure 9 shows significant main effects for average speed of rotation. The error bars denote 95% confidence intervals.

Table 5. Means for average rotation speed for cell phone speech at levels 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs

Level					
0 dB	2 dB	4 dB	6 dB	8 dB	10 dB
11.660	10.931	10.399	9.865	9.243	9.849

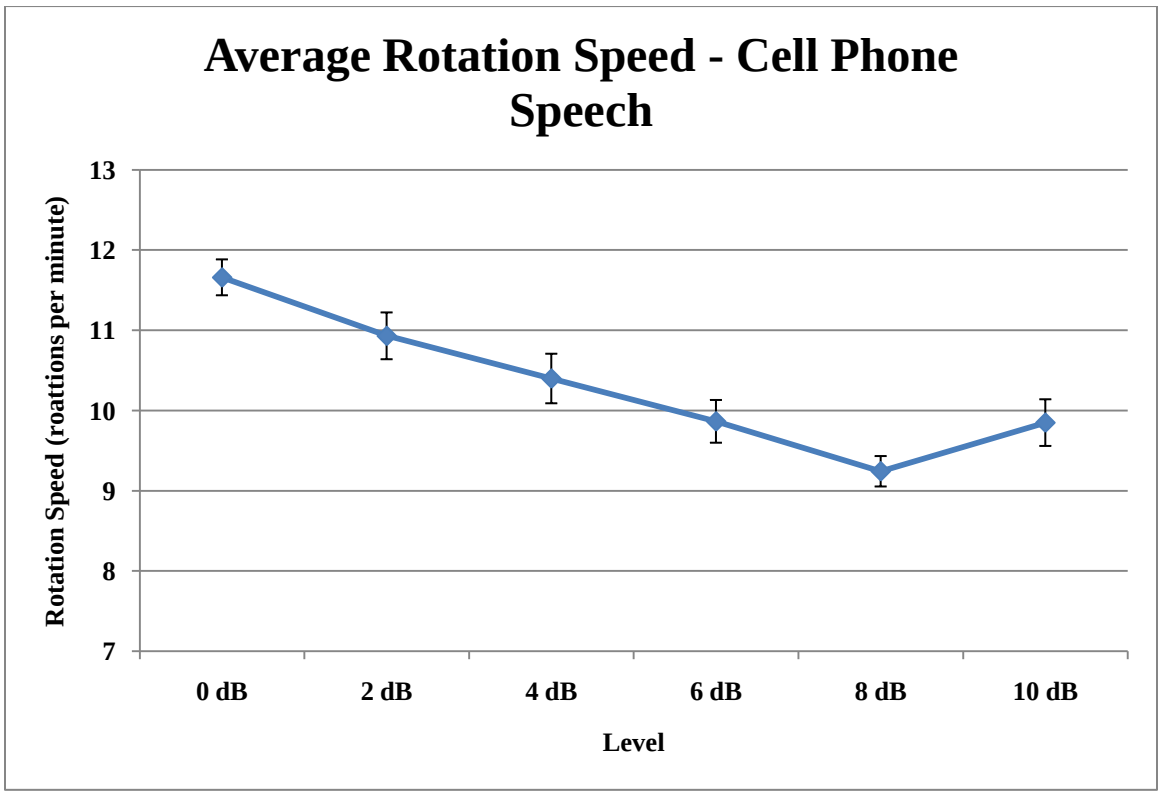

Figure 9. Overall mean rotation speed at levels 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs – cell phone speech

Rotation Speed - Natural Speech

A 2-way within groups ANOVA was used to examine the main effects and interactions of level (-8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs) as they relate to the average rotation speed in the adaptive pursuit rotor task when natural phone speech was presented to the participants. There was a main effect of level ($F(5, 85) = 4.0$, $p = 0.003$, $M_{se} = 1.741$). The pattern of mean differences was that at low SNR, there was no significant difference in average speed of rotation between successive SNRs ($LSD_{mmd} = 0.622$). However, there was significant difference between average speed of rotation between 0 dB SNR and 2 dB SNR. Table 6 shows the average speed of rotation during at levels -8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs. Figure 10 shows significant main effects for average speed of rotation. The error bars denote 95% confidence intervals.

Table 6. Means for average rotation speed for natural speech at levels -8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs

Level					
-8 dB	-6 dB	-4 dB	-2 dB	0 dB	2 dB
8.285	8.451944	8.489444	8.623056	8.848889	9.514444

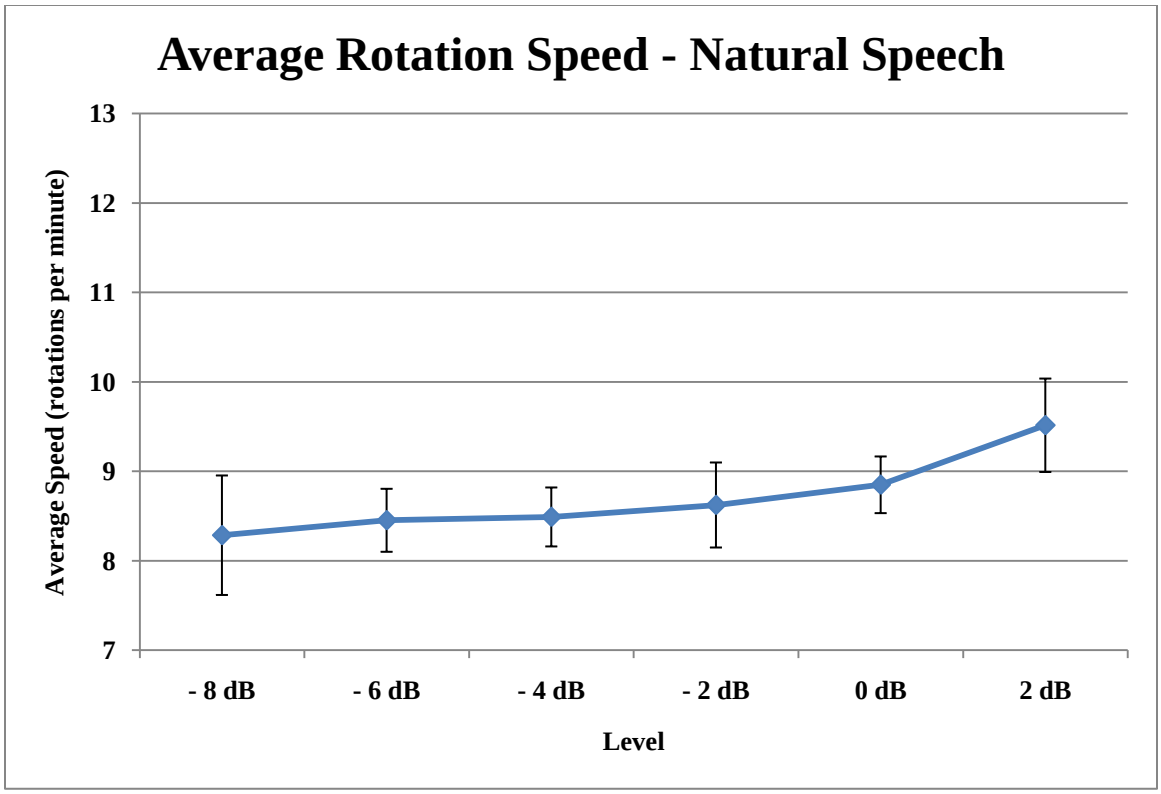

Figure 10. Overall mean rotation speed at levels -8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs – natural speech

The next research question investigated the relationship between the signal-to-noise ratios of the transmitted speech stimuli and the performance in a word repetition task. To answer this question, word accuracy on auditory word repetition task was compared at the different levels of stimuli.

Word Accuracy - Cell Phone Speech

A 2-way within groups ANOVA was used to examine the main effects and interactions of level (0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs as they relate to word accuracy in the auditory word repetition task. There was a main effect of level of speech ($F(5, 85) = 289.6$, $p < 0.001$, $M_{se} = 71.084$).

The pattern of mean differences was that as the signal got better, the performance on auditory word repetition task improved (LSD_{mmd} = 3.974). Also, the successive differences between the levels decreased as the signal got better. Table 7 shows the average word accuracy for predictable and unpredictable sentences and at levels 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs. Figure 11 shows the significant main effects for word accuracy. The error bars denote 95% confidence intervals.

Table 7. Means for word accuracy for cell phone speech at levels 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs

Level					
0 dB	2 dB	4 dB	6 dB	8 dB	10 dB
13.61111	31.38889	47.91667	59.86111	69.58333	76.52778

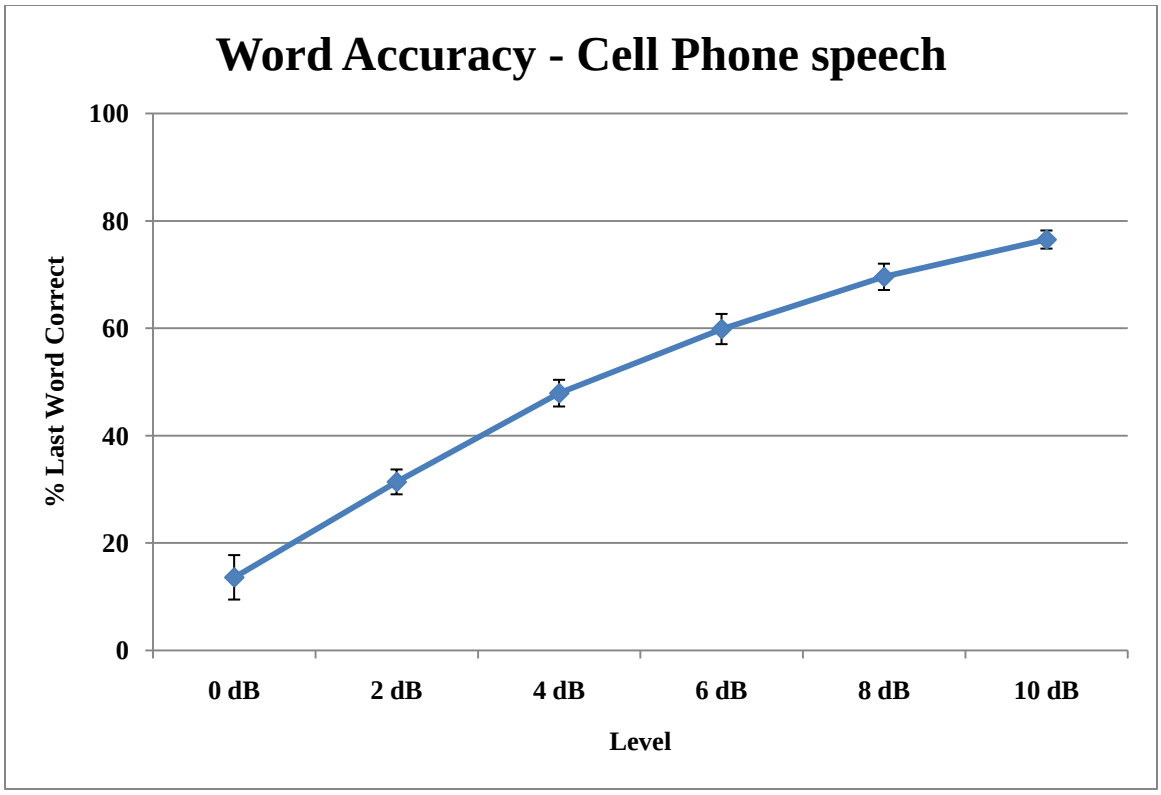

Figure 11. Overall means for word accuracy – cell phone speech

Word Accuracy - Natural Speech

A 2-way within groups ANOVA was used to examine the main effects and interactions of level (-8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs) as they relate to word accuracy in the auditory word repetition task. There was a main effect of level of speech ($F(5, 85) = 702.2$, $p < 0.001$, $M_{se} = 61.684$). The pattern of mean differences was that as the signal got better, the performance on auditory word repetition task improved ($LSD_{mmd} = 3.702$). Table 8 shows the average word accuracy for predictable and unpredictable sentences and at levels -8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs. Figure 12 shows the significant main effects for word accuracy. The error bars denote 95% confidence intervals.

Table 8. Means for word accuracy for natural speech at levels -8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs

Level					
-8 dB	-6 dB	-4 dB	-2 dB	0 dB	2 dB
4.305556	20.27778	59.72222	74.30556	84.02778	86.66667

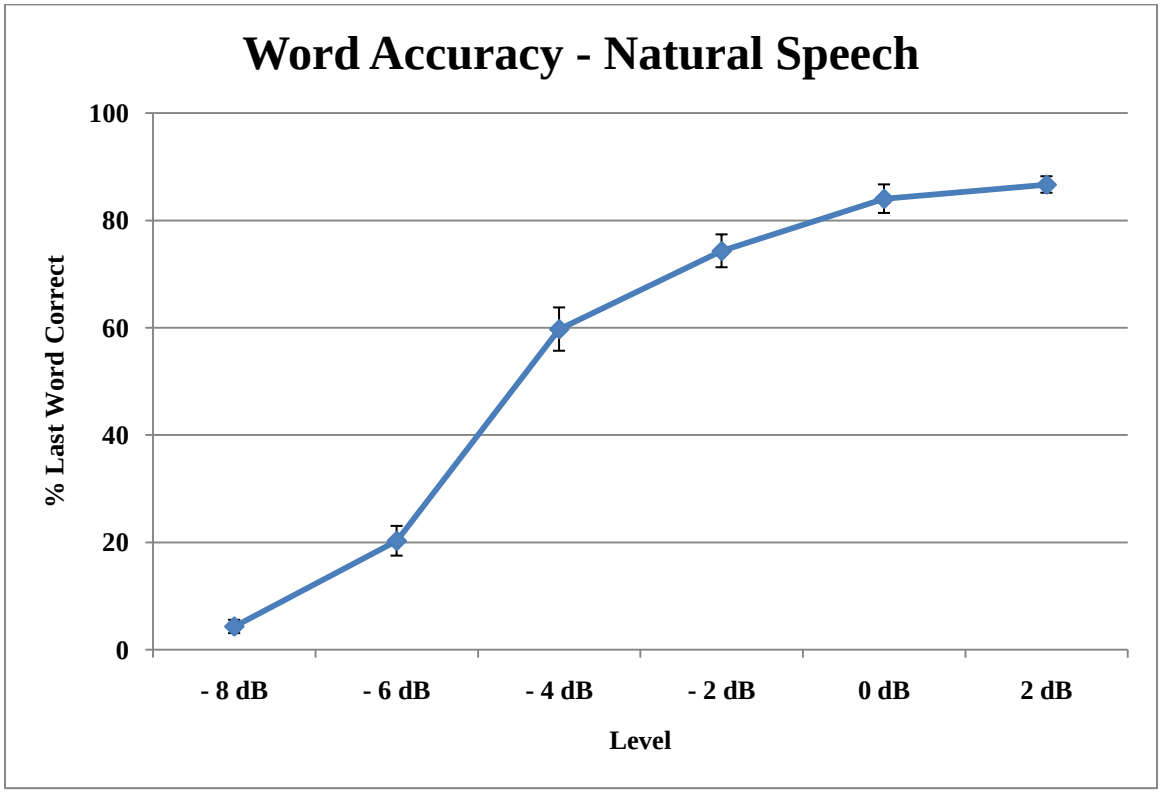

Figure 12. Overall means for word accuracy – natural speech

The next research question investigated the effect of semantic context of the presented sentences on performance on simultaneous visual-motor task and auditory word repetition task. To answer the first part of the research question, the average rotation speed on adaptive pursuit rotor task was compared for predictable and unpredictable sentences. To answer the second

part of the research question, word accuracy on word repetition task was compared for predictable and unpredictable sentences.

A 2-way within groups ANOVA was used to examine the main effects and interactions of context (predictable and unpredictable) as they relate to the average rotation speed in the adaptive pursuit rotor task when cell phone speech and natural speech was presented to the participants. For cell phone speech, there was a main effect of context ($F(1, 17) = 289.6$, $p < 0.001$) with unpredictable sentences having higher speed of rotation than predictable sentences. However, for natural speech, there was no effect of context on the performance of the adaptive pursuit rotor task. Table 9 shows the average speed of rotation for predictable and unpredictable sentences when cell phone speech and natural speech was presented to the participants. Figure 13 shows effect of semantic context on simultaneous visual-motor task for cell phone and natural speech. The error bars denote 95% confidence intervals.

Table 9. Means for average rotation speed for predictable and unpredictable sentences for cell phone speech and natural speech

Cell Phone Speech		Natural Speech	
Predictable	Unpredictable	Predictable	Unpredictable
9.662	10.988	8.701759	8.7025

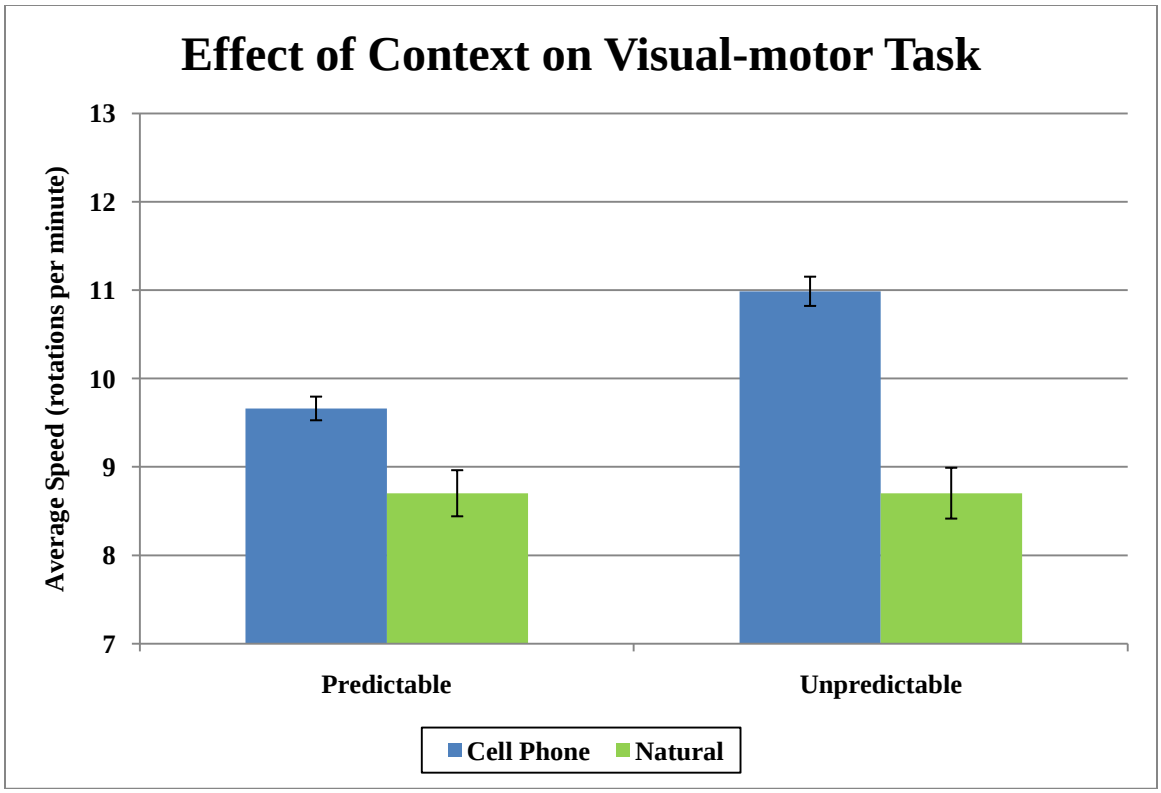

Figure 13. Effect of context on simultaneous visual-motor task for cell phone and natural speech

A 2-way within groups ANOVA was used to examine the main effects and interactions of context (predictable and unpredictable) as they relate to the word accuracy during auditory word repetition task when cell phone speech and natural speech was presented to the participants. For cell phone speech, there was a main effect of semantic context of speech ($F(1, 17) = 1638.8$, $p < 0.001$) with higher word accuracy for predictable sentences (66%) as compared to unpredictable sentences (33.6%). For natural speech, there was a main effect of context of speech ($F(1, 17) = 535.9$, $p < 0.001$) with higher word accuracy for predictable sentences (66.3%) as compared to

unpredictable sentences (43.5%). Though there was no difference between the cell phone and speech and natural speech in word accuracy for predictable sentences. However, for unpredictable sentences, natural speech had a higher word accuracy compared to cell phone speech. On the whole, context of the sentence helped cell phone speech a lot compared to natural speech. Table 10 shows the percent last words correct for predictable and unpredictable sentences when cell phone speech and natural speech was presented to the participants. Figure 14 shows effect of semantic context on auditory word repetition task for cell phone and natural speech. The error bars denote 95% confidence intervals.

Table 10. Means for word accuracy for predictable and unpredictable sentences for cell phone speech and natural speech

Cell Phone Speech		Natural Speech	
Predictable	Unpredictable	Predictable	Unpredictable
66	33.6	66.3	43.5

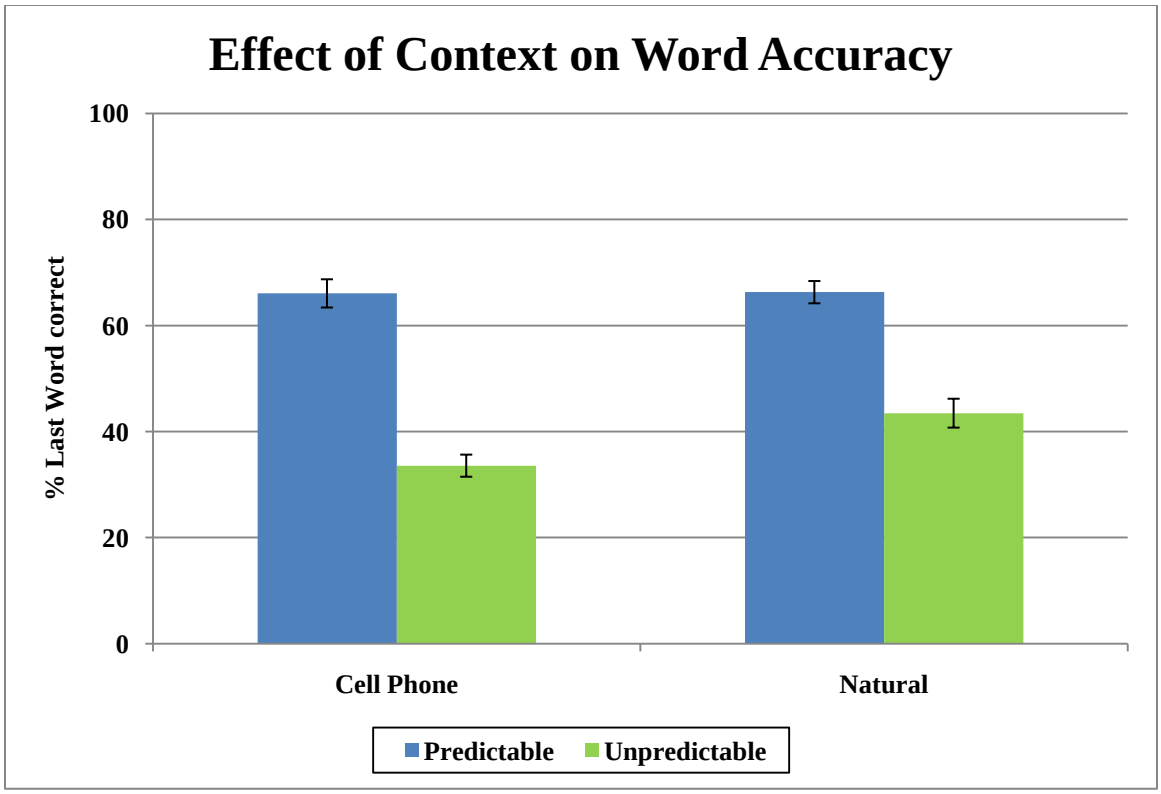

Figure 14. Effect of context on auditory word repetition task for cell phone and natural speech

The next research question investigated the cognitive load imposed on the listener due to listening to cell phone speech and natural speech. To answer the research question, the average NASA-TLX score reported by the user was compared for different levels of the signal presented.

A 2-way within groups ANOVA was used to examine the main effects and interactions of cell phone speech presented at different levels (0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs as they relate to NASA task load index (NASA-TLX). There was a main effect of level of speech ($F(5, 85) = 10.7$, $p < 0.001$, $M_{se} = 300.933$). The pattern of mean differences was that at very low

SNRs, there was no significant difference in the task load index between successive SNR levels (LSD_{mmd} = 8.18). However, the task load index at10 dB SNR was significantly lower than the task load index at 8 dB SNR. Table 11 shows the average task load index for cell phone speech at levels 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs. Figure 15 shows the significant main effects for average task load index. The error bars denote 95% confidence intervals.

Table 11. Means for average task load index for cell phone speech at levels 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs

Level					
0 dB	2 dB	4 dB	6 dB	8 dB	10 dB
81.7875	76.08361	68.77917	63.54611	64.96333	55.29556

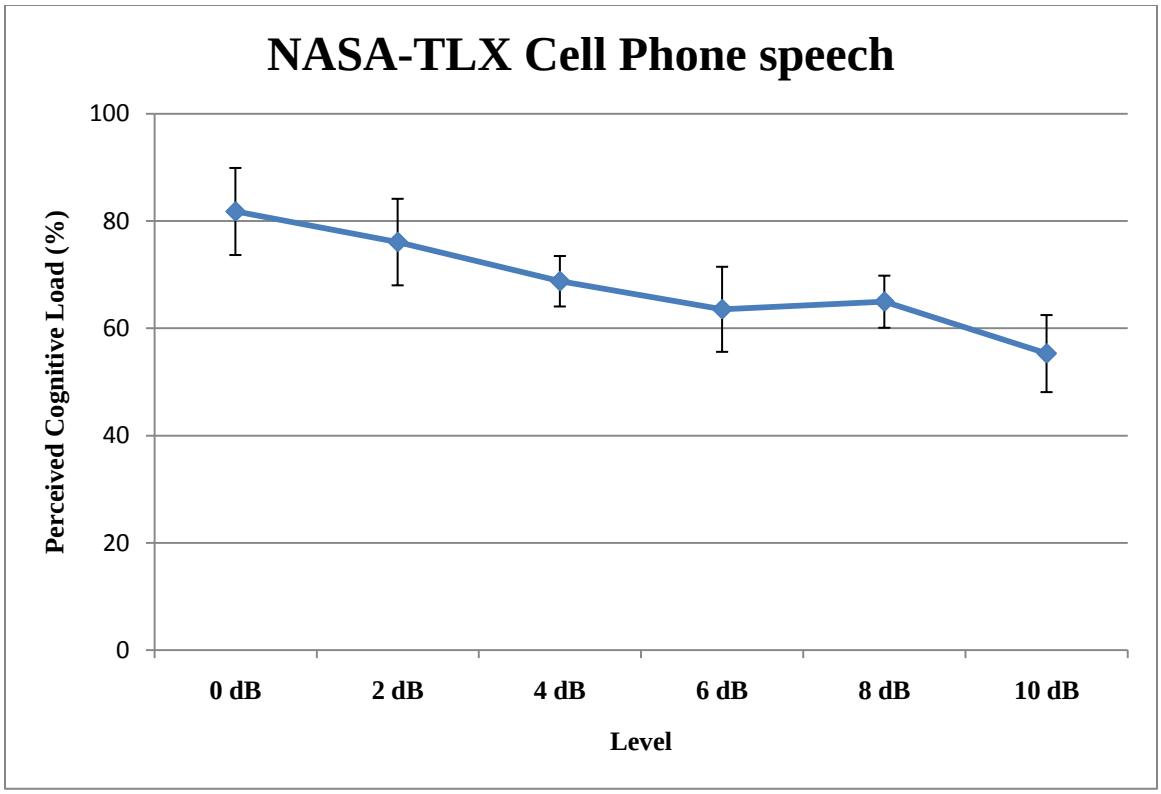

Figure 15. Overall mean cognitive load at levels 0 dB, 2 dB, 4dB, 6 dB, 8 dB, and 10 dB SNRs – cell phone speech

A 2-way within groups ANOVA was used to examine the main effects and interactions of natural speech presented at different levels (-8 dB, -6 dB, - 4 dB,-2dB, 0 dB, and 2 dB SNRs as they relate to NASA task load index (NASA-TLX). There was a main effect of level of speech ($F(5, 85) = 24.6$, $p < 0.001$, $M_{se} = 363.916$). The pattern of mean differences was that very low SNRs, there was significant difference in the task load index between -6 dB SNR and -4 dB SNR ($LSD_{mmd} = 8.992$). Also, there was a significant difference in the task load index between -2 dB SNR and 0 dB SNR. Table 12 shows the average task load index for natural speech at levels -8 dB, -6 dB, -4dB, -2 dB,

0 dB, and 2 dB SNRs. Figure 16 shows the significant main effects for average task load index. The error bars denote 95% confidence intervals.

Table 12. Means for average task load index for natural speech at levels -8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs

Level					
-8 dB	-6 dB	-4 dB	-2 dB	0 dB	2 dB
91.62972	82.92611	68.96333	66.71583	57.28722	49.17556

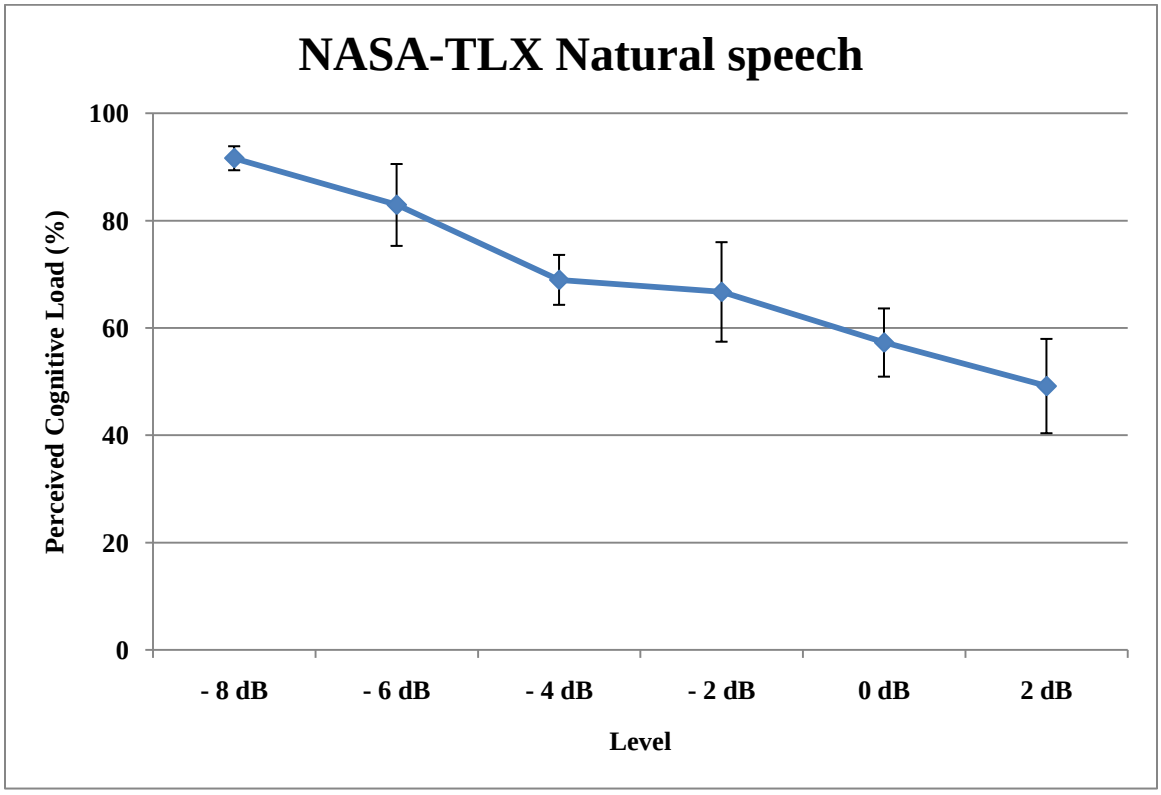

Figure 16. Overall mean cognitive load at levels -8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB SNRs – natural speech

For cell phone speech, there was a main effect of context of speech ($F(1, 17) = 69.4$, $p < 0.001$) as they relate to the average NASA-TLX score reported by the participants. The task load for unpredictable sentences (75.10) was higher as compared to predictable sentences (61.72). For natural speech, there was a main effect of context of speech ($F(1, 17) = 44.8$, $p < 0.001$) as they relate to the average NASA-TLX score reported by the participants. The task load for unpredictable sentences (73.91) was higher as compared to predictable sentences (64.99). The difference in the average NASA-TLX score reported by the participants between predictable and unpredictable sentences was higher for cell phone speech as compared to natural speech. Table 13 shows the average NASA-TLX score reported by the participants for predictable and unpredictable sentences when cell phone speech and natural speech was presented. Figure 17 shows effect of semantic context on the average NASA-TLX score reported by the participants for cell phone and natural speech. The error bars denote 95% confidence intervals.

Table 13. Means for the average task load index for predictable and unpredictable sentences for cell phone speech and natural speech

Cell Phone Speech		Natural Speech	
Predictable	Unpredictable	Predictable	Unpredictable
61.72	75.1	64.99	73.91

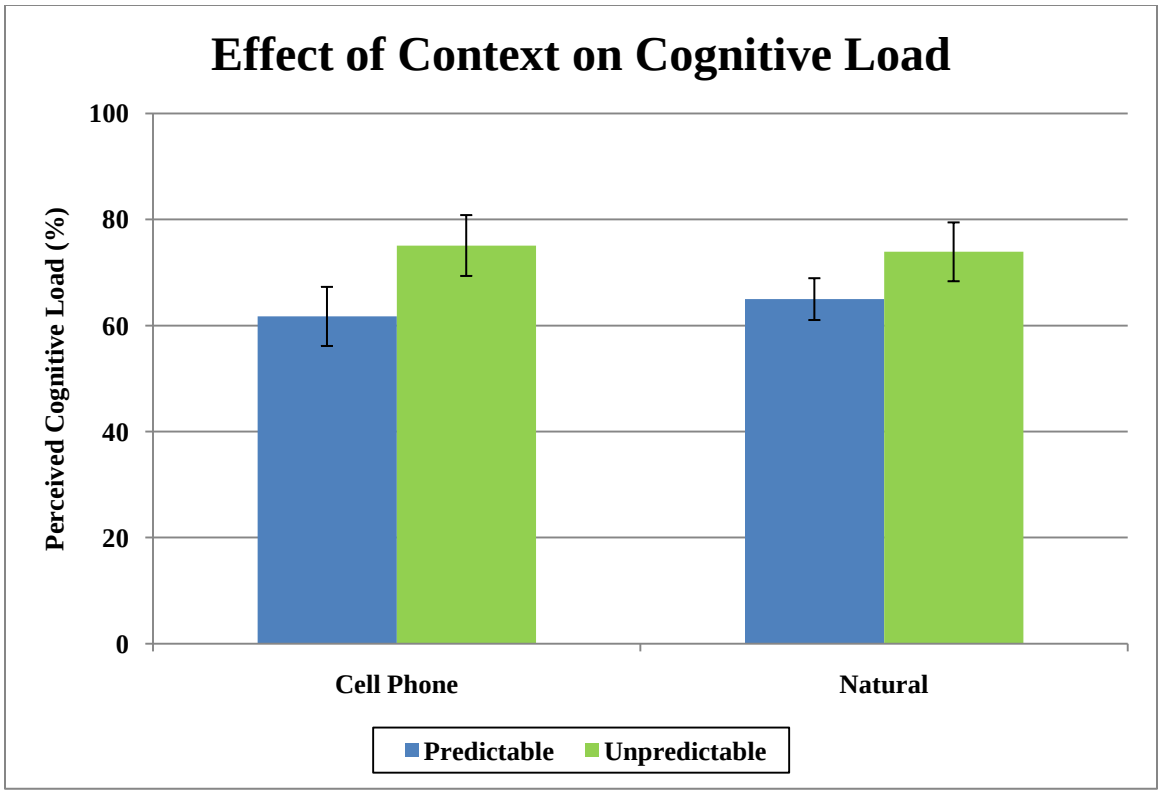

Figure 17. Effect of context on the average NASA-TLX score reported by the participants for cell phone and natural speech

The last research question investigated the effects of semantic context at different stimulus presentation levels for cell phone and natural speech in an auditory word repletion task. To answer this question, word accuracy on word repetition task was compared at the different levels of stimuli during predictable and unpredictable sentences for cell phone and natural speech.

For cell phone speech, there was a significant 2-way interaction between level and context as they relate to word accuracy ($F(5, 85) = 52.3$, $p < 0.001$, $M_{se} = 41.569$). Visual Inspection of the cell means (using $LSD_{mmd} = 4.298$) revealed that as the quality of the signal improved, their performance

on auditory word repetition task improved. This was true for both predictable and unpredictable sentences. However, predictable last word sentences showed a higher change in word accuracy as compared to unpredictable sentences. Table 14 shows the word accuracy for predictable and unpredictable sentences at levels 0 dB, 2 dB, 4 dB, 6 dB, 8 dB, and 10 dB SNRs. Figure 18 shows the interaction of level by context for cell phone speech. The error bars denote 95% confidence intervals.

Table 14. Means for word accuracy (%) for predictable and unpredictable sentences at levels 0 dB, 2 dB, 4 dB, 6 dB, 8 dB, and 10 dB SNRs

	0 dB	2 dB	4 dB	6 dB	8 dB	10 dB
Predictable	15.83	45	70.28	82.5	90.28	92.5
Unpredictable	11.39	17.78	25.56	37.22	48.89	60.56

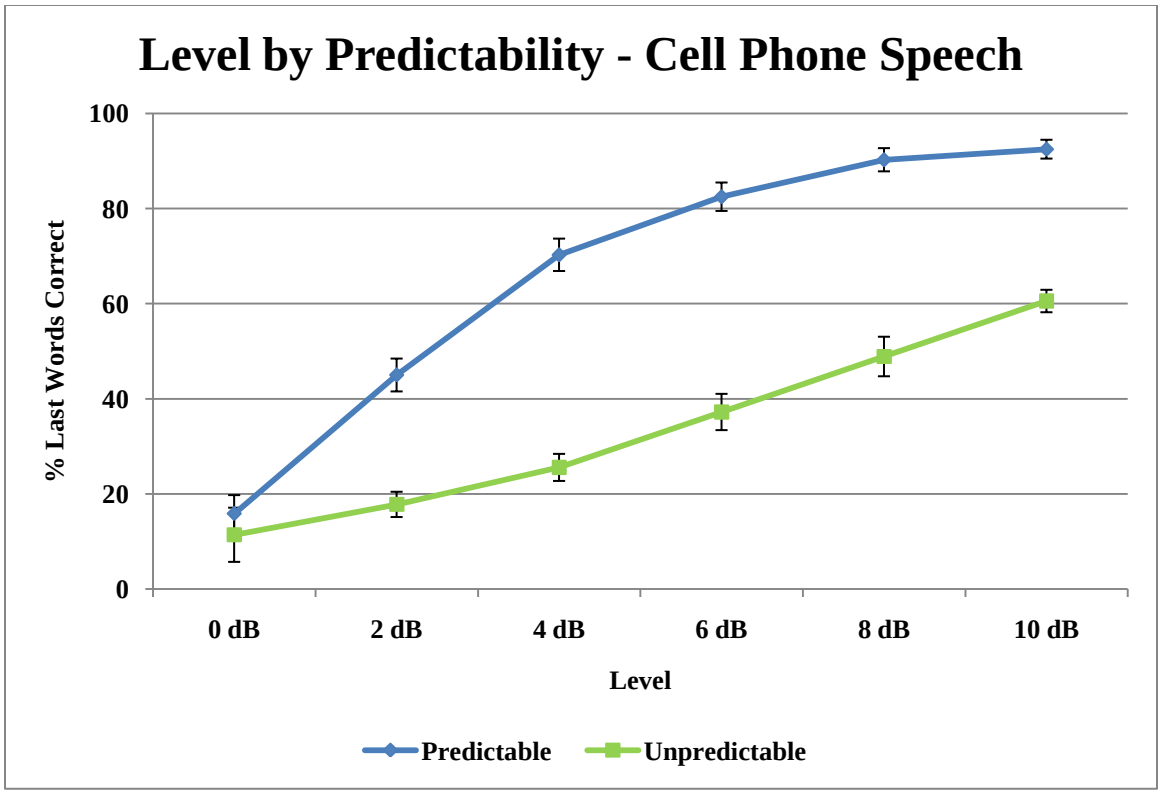

Figure 18. Level by context interaction for cell phone speech

Other Results.

For natural speech, there was a significant 2-way interaction between level and context as they relate to word accuracy ($F(5, 85) = 38.6$, $p < 0.001$, $M_{se} = 46.848$). Visual Inspection of the cell means (using $LSD_{mmd} = 4.139$) revealed that as the quality of the signal improved, the performance on auditory word repetition task improved. This was true for both predictable and unpredictable sentences. However, predictable last word sentences showed a higher change in word accuracy as compared to unpredictable sentences. Table 15 shows the word accuracy for predictable and unpredictable sentences at levels -8 dB, -6 dB, -4dB, -2 dB, 0 dB, and 2 dB

SNRs. Figure 19 shows the interaction of level by context for natural speech. The error bars denote 95% confidence intervals.

Word accuracy was higher for predictable sentences as compared to unpredictable sentences for both cell phone speech and natural speech. However, the difference in percent last word between predictable and unpredictable sentences for successive levels of SNR was higher for cell phone speech as compared to natural speech. It should be noted that cell phone speech and natural speech were presented in totally different SNRs.

Table 15. Means for word accuracy for predictable and unpredictable sentences at levels -8 dB, -6 dB, - 4 dB, -2dB, 0 dB, and 2 dB SNRs

	-8 dB	-6 dB	-4 dB	-2 dB	0 dB	2 dB
Predictable	6.11	25	80.56	90.28	97.5	98.33
Unpredictable	2.5	15.56	38.89	58.33	70.56	75

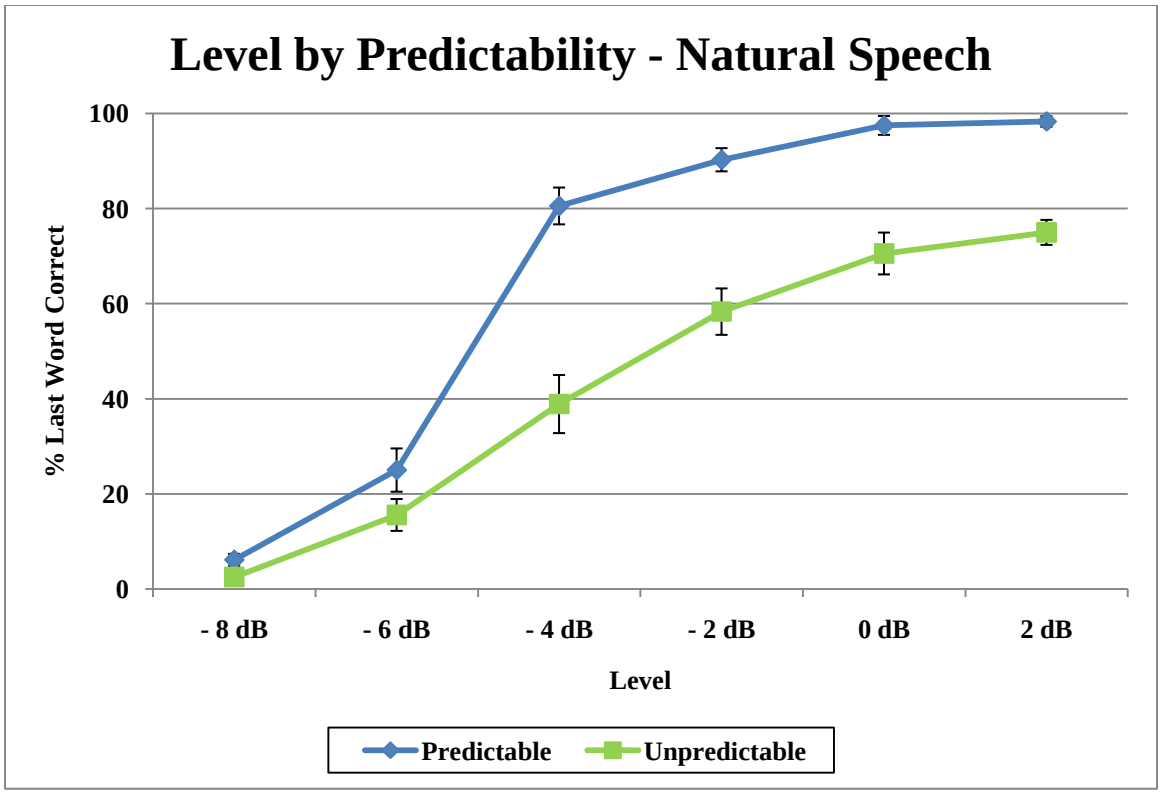

Figure 19. Level by context interaction for natural speech

Comprehensive findings.

It was also found that, at low intelligibility levels, cell phone speech had a higher speed of rotation than natural speech. Figure 20 shows the relationship between intelligibility and visual motor performance and cognitive load and visual motor performance for predictable sentences. The participants had to spend more resources for the auditory task while listening to natural speech as compared to cell phone speech. However, it should be noted that there was an 8 dB difference in SNRs between cell phone speech and natural speech at these intelligibility levels. The difference in APR performance while listening to natural and cell phone speech at lower

intelligibility levels proves that cell phone and natural speech were processed differently at these intelligibility levels. As the intelligibility of the sentences presented increased, the performance on APR task while listening to natural speech started to increase whereas it started to decrease for cell phone speech. This showed that as the intelligibility increased, the participants spent fewer resources to the auditory task while listening to natural speech and more resources to the auditory task while listening to cell phone speech. After certain intelligibility threshold, the intelligibility level of cell phone signals was high so that the participants could spend more resources to the tracking task there by there was an increase in speed of rotation.

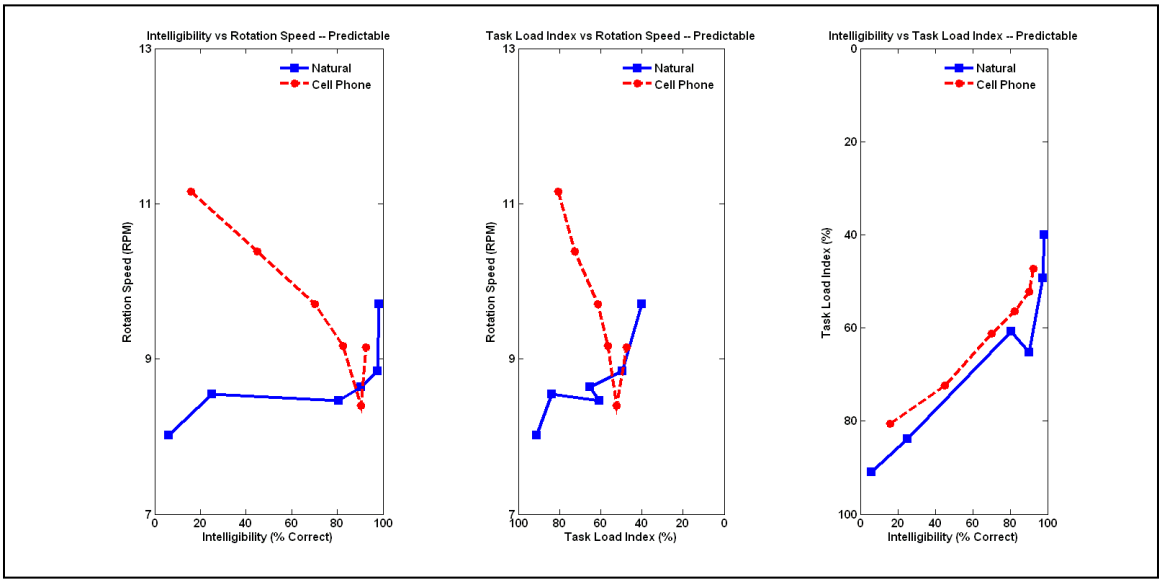

Figure 20. Relationship between intelligibility and visual motor performance, cognitive load and visual motor performance, and intelligibility and cognitive load for predictable sentences

It has been found that the participants give up while listening to speech at lower intelligibility levels (up to 80 % intelligibility) (Beukelman, D. B., personal communication, February, 2010). Figure 21 shows the relationship between intelligibility and visual motor performance for predictable sentences for intelligibility regions greater than 75%. It should be noted that both the axes in figure 21 are dependent measures from experiment 2. For natural speech, as intelligibility increased, rotation speed increased. As intelligibility of the stimuli presented increased, the participants started spending fewer resources to the auditory word repetition task and more resources adaptive tracking task. At these intelligibility levels, the presented signal was so highly intelligible that the participants achieved better results by spending fewer resources. The excess resources were spent to the adaptive tracking task and hence average speed increased. At very high intelligibility levels (> 95%), small increase in intelligibility led to a big increase in average speed.

For cell phone speech, as intelligibility increased, rotation speed started to decrease. After an intelligibility threshold, rotation speed started to increase. It could be concluded that, at lower intelligible levels, the signal was so less intelligible that the participants allocated more resources to the adaptive tracking task and less resources to the auditory word repetition task. As intelligibility increased, the participants started allocating more resources to the auditory word repetition task and fewer resources to the

adaptive tracking task. After achieving the threshold, the stimuli presented was so highly intelligible that the participants spent less resources to the auditory word repetition task and achieved greater results. The additional resources were spent to the adaptive tracking task and hence average speed increased.

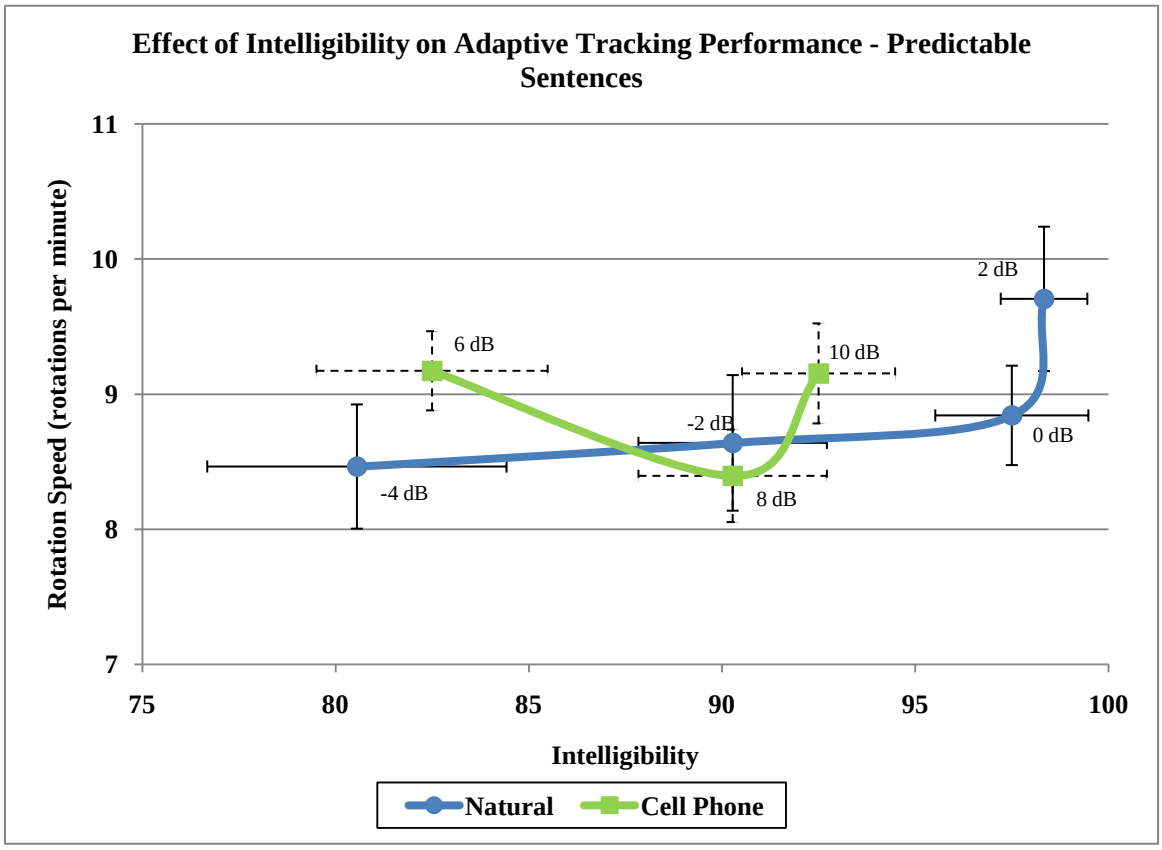

Figure 21. Relationship between intelligibility and visual motor performance for predictable sentences for intelligibility regions > 75%

The participants allocated more resources to listening when the auditory stimulus presented was natural speech than when the signal was cell phone speech. However, it should be noted that there was an 8 dB

difference in SNRs between cell phone speech and natural speech at these intelligibility levels. It has been found in the literature that the participants give up while listening to speech at lower intelligibility levels (up to 80 % intelligibility). The difference in APR performance while listening to natural and cell phone speech at lower intelligibility levels suggests that cell phone and natural speech were processed differently at these intelligibility levels. As the intelligibility of the sentences presented increased, the performance on APR task while listening to natural speech started to increase whereas it started to decrease for cell phone speech. This showed that as the intelligibility increased, the participants spent fewer resources to the auditory task while listening to natural speech and more resources to the auditory task while listening to cell phone speech. After certain intelligibility threshold, the intelligibility level of cell phone signals was high so that the participants could spend more resources to the tracking task there by there was an increase in speed of rotation.

When NASA-TLX, APR performance, and speech intelligibility were compared, it was found that for cell phone speech, at low intelligibility levels, the participants had a higher cognitive load and higher speed of rotation. As intelligibility increased, both average speed and average cognitive load started to decrease. However, after a threshold at intelligibility, both average speed and average cognitive load started to increased. For natural speech, at low intelligibility levels, the participants had a higher cognitive load and

lower speed of rotation. As intelligibility increased, average cognitive load started to decrease but there was no difference in average speed of rotation. However, at higher intelligibility levels, the average speed of rotation increased and average cognitive load decreased.

There was a high positive correlation between average speed of rotation and intelligibility for natural speech indicating that increased intelligibility increased the performance on the tracking task ($r = 0.304$, $p = 0.001$). However, the correlation between average speed of rotation and intelligibility for cell phone speech was negative indicating that increased intelligibility decreased the performance on the tracking task ($r = -0.717$, $p < 0.001$). This indicated that while listening to cell phone speech, at lower intelligibility levels, the participants spent more resources to the tracking task and probably neglected the auditory task. As the intelligibility increased, the participants paid more attention to the auditory task and less attention to the visual task. However, for natural speech, the participants paid more attention to the auditory task and less attention to the visual task even though the intelligibility of the presented sentences was low. As intelligibility increased, the participants started paying more attention to the visual task.

There was a high negative correlation between average speed of rotation and average cognitive load measured using NASA-TLX for natural speech indicated that increased speed led to decreased cognitive load ($r = -0.3$,

p = 0.002). However, there was a high positive correlation between average speed of rotation and average cognitive load measured using NASA-TLX for cell phone speech indicated that increased speed led to increased cognitive load ($r = 0.415$, $p < 0.001$).

There was a high negative correlation between intelligibility and average cognitive load measured using NASA-TLX for natural speech indicated that increased intelligibility led to decreased cognitive load ($r = -0.724$, $p < 0.001$). This was true for cell phone speech too ($r = -0.568$, $p < 0.001$).

When NASA-TLX, APR performance, and speech intelligibility for unpredictable sentences were compared, it was found that for cell phone speech, at low intelligibility levels, the participants had a higher cognitive load and higher speed of rotation. As intelligibility was nearing threshold, there was a sudden increase in average cognitive load. However, after a threshold at intelligibility, both average speed and average cognitive load started to increase. For natural speech, at low intelligibility levels, the participants had a higher cognitive load and lower speed of rotation. As intelligibility increased, average cognitive load started to decrease but there was no difference in average speed of rotation. However, at higher intelligibility levels, the average speed of rotation increased and average cognitive load decreased. Figure 22 shows the relationship between

intelligibility and visual motor performance and cognitive load and visual motor performance for unpredictable sentences.

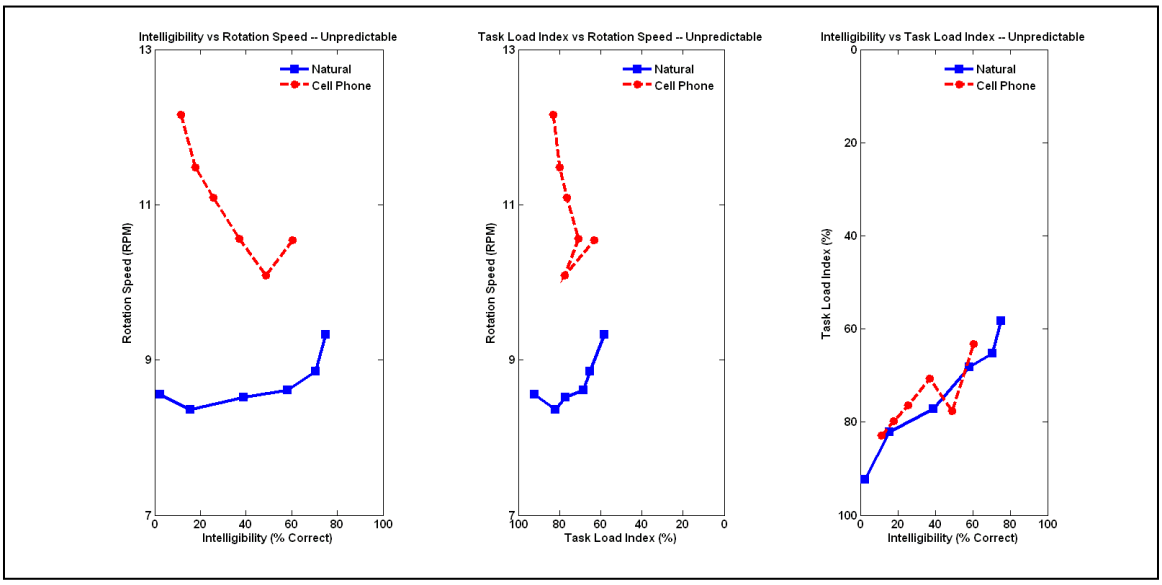

Figure 22. Relationship between intelligibility and visual motor performance, cognitive load and visual motor performance, and intelligibility and cognitive load for unpredictable sentences

There was no significant correlation between average speed of rotation and intelligibility for natural speech indicating that increased intelligibility did not have any effect on the performance on the tracking task. However, the correlation between average speed of rotation and intelligibility for cell phone speech was negative ($r = -0.547$, $p < 0.001$) indicating that increased intelligibility decreased the performance on the tracking task. This indicated that while listening to cell phone speech, at lower intelligibility levels, the participants spent more resources to the tracking task and probably neglected the auditory task. As the intelligibility increased, the participants paid more attention to the auditory task and less attention to the visual task.

However, for natural speech, the participants paid equal attention to the auditory task and the visual task even though the intelligibility of the presented sentences was low. As intelligibility increased, there was no change in the manner in attentional resources were allocated.

There was a no significant correlation between average speed of rotation and average cognitive load measured using NASA-TLX for natural speech and cell phone speech. This indicated that increased speed did not have any effect on perceived cognitive load for both natural speech and cell phone speech. There was a low negative correlation between intelligibility and average cognitive load measured using NASA-TLX for cell phone speech indicated that increased speed led to decreased cognitive load ($r = -0.262$, $p = 0.006$). However, there was a high negative correlation between intelligibility and average cognitive load measured using NASA-TLX for natural speech indicated that increased speed led to increased cognitive load ($r = -0.579$, $p < 0.001$). Based on these correlations, it could be concluded that average speed had a higher effect on perceived cognitive load for natural speech as compared to cell phone speech.

Overall, these results showed how speech quality factors interact with intelligibility, semantic context, and simultaneous visual-motor performance. However, they do not address the role of attention in these performance measures. Hence, a different experiment was designed to look into the role of attention in the dependent measures used in experiment 2.

CHAPTER VI

Experiment 3

There is substantial need for understanding speech perception and its underlying processes to understand the influence of speech on driving and other simultaneous tasks and the influence of driving and other visual-motor tasks on speech perception. Driving has been described as predominantly an automatic task which requires conscious effort only occasionally in demanding situations such as driving in a new town (Tarawneh, 1991). On the other hand, talking on a cell phone has been categorized as a controlled processing task (Matthews et al., 2003). Unfortunately, most theories of speech perception were developed with limited demands on attention and very few systematically-varied attentional parameters. Therefore, more complete models of speech perception must be developed that consider attention as an integral component of perception.

Single resource theory (Kahneman, 1973) argued the existence of a single pool of resources with limited capacity for performing variety of tasks. Also, the amount of resources allocated for the performance different tasks depends upon the difficulty of the individual tasks. The participant devises an allocation strategy for dividing the available limited resources for various tasks based on the characteristics of the stimuli and individual motivation.

Multiple resource theory (Wickens, 1984) was an alternative theory to single resource theory. MRT has an explanation for the good performance

of trained participants in tasks that required divided attention. According to MRT, humans have several different pools of resources that could be accessed simultaneously for the performance of multiple tasks. The nature and difficulty of the tasks dictated whether the resources were drawn from the same pool of resources or different pool of resources and whether the tasks were completed in sequential or in parallel. The tasks were performed in sequential manner if the tasks require accessing the same pool of resources. However, the tasks were performed in parallel if the tasks require accessing different pool of resources. The original model specified only audition and vision as modalities, in principle; the model would apply to any other modality as well.

Most of the early resource theories were developed based on the following factor: available resources were allocated and spent for the performance of different tasks based on the difficulty of task. Of all the theories as explained in introduction, the most appropriate model of attention for the current experiment was the model proposed by Schneider & Shiffrin's automatic versus controlled processing model (1977).

Traditionally, two major factors have been proposed to increase the chances of being involved in an accident while driving and simultaneously using a cell phone. One factor is the visual and mechanical competition between driving and using a cell phone. While placing or receiving calls, the drivers may momentarily remove their vision from the road and at least one

hand from the steering wheel and look at the cell phone for operating it. The second factor is the cognitive competition between driving and operating a cell phone. A person's ability to concurrently do two or more tasks is generally limited to one task requiring conscious effort (controlled processing) and one or more tasks requiring little to no conscious effort (automated processing) (Schneider et al., 1984). That is, only one task requiring controlled processing can be performed at once without influencing one of the other tasks. Therefore, it is important to refine our knowledge regarding the extent to which and under what circumstances listening is an automated versus controlled task.

Experiment 3 was designed to study the effects of automatic and controlled processing tasks on cell phone speech perception and simultaneous visual-motor tracking performance. Consistent Mapping (CM) and Varied Mapping (VM) tasks were used to experimentally induce automatic and controlled processing. Experiment 3 was designed to answer the following research questions.

1. Are different speech sources (natural versus cell phone sentences) processed differently during simultaneous consistent mapping and varied mapping tasks?

2. Do different speech source qualities require different attention mechanisms?

3. Are speech and visual-motor tasks affected differentially by consistent mapping versus varied mapping tasks?

4. How is the simultaneous visual-motor task affected by adding additional visual word recognition and memory tasks?

5. What effect does CM and VM tasks have on the performance on visual word identification task while listening to cell phone and natural speech?

Participants

Sixteen participants participated in Experiment 3. All participants were listeners with normal hearing and eyesight. All spoke English as their primary language. The participants were female undergraduate students in the Special Education and Communication Disorders Department of the University of Nebraska – Lincoln. They ranged in age from 20 to 28 with a mean age of 21.25. The participants were randomly assigned to either the consistent mapping identification group or varied mapping identification group.

Experimental Tasks

There were four conditions and six sessions of approximately one hour each in the entire experiment. The first four sessions were the same for all the participants. The first condition, single-task condition, lasted for the first two sessions. During the single-task condition, the participants performed

visual word recognition task alone with their non-dominant hand. The second condition, dual-task condition, lasted for the third and fourth sessions. During the dual-task condition, the participants performed visual word recognition task using their non-dominant hand and adaptive pursuit tracking task using their dominant hand. Sessions five and six were the triple-task condition of the study. In the triple-task condition, the effect of mapping was systematically varied. Participants in the consistent mapping group performed consistent mapping first followed by varied mapping whereas the participants in the varied mapping group performed varied mapping followed by consistent mapping. During the triple-task condition, the participants performed three tasks simultaneously: Visual word identification with their non-dominant hand, adaptive tracking with their dominant hand and responding to auditory stimuli presented via the headphone. The participants stop performing the visual word recognition and adaptive tracking tasks when all the stimuli in a given condition (20 sentences per condition) were presented. The target words during the consistent mapping condition were the same target words used during the sessions 1 to 4. The target words during the varied mapping condition changes for every block of trials.

Stimuli and Apparatus

Adaptive pursuit rotor task.

Participants performed an adaptive pursuit tracking task as the simultaneous task with speech perception. In the adaptive pursuit rotor task, the participants used a stylus to move a cursor on a computer display to keep it aligned as closely as possible to a moving target. This task was identical to that described in Experiment 1.

Auditory word repetition task.

In auditory word repetition task, the participants were asked to repeat the last word of the sentence which they just heard. This task was identical to that described in Experiment 1.

Visual word recognition task.

The visual word recognition task required the participants to respond to the visual stimuli presented using their non-dominant hand. Target words were presented to the participants before the start of the experiment proper. Four words arranged in a 2X2 matrix were presented to the participants. The visual stimuli were presented on a second 15" computer screen situated at about 20 degrees to the left of the participant. The participants were asked to press "Present" button if one of the four target words were present in the visual stimuli. If none of the four target words were present, the participants were asked to press "Absent" button.

Stimuli

The stimuli for the visual word recognition task consisted of 100 four letter words. The words were selected from the 2,938 monosyllabic words that were rated for the subjective frequency by young and older adults using 7-point scale (Balota et al., 2001). For individual words, see Appendix 2.

The stimuli for auditory word repetition consisted of cell phone speech missed with multitalker babble at 4 dB, 6 dB, and 8 dB signal-to-noise ratios and natural speech mixed with multitalker babble (Bilger, R. C., et al, 1984) at 0 dB, -2 dB and -4 dB signal-to-noise ratios. The sentences were presented to the participants via circumaural headphones at approximately 68 dB SPL. The levels for cell phone speech and natural speech were selected to have similar intelligibility levels.

Construction of the visual stimuli.

The stimuli for the lexical decision task were created using CorelDraw X4. The page was divided into a 2 X 2 matrix and words were typed using 52 points with an Arial typeface. 200 sets of four words were selected from the pool of 100 words without replacement. 80 word sets were randomly selected and one of the four words in each of the word set was replaced with one of the four target words. After replacement, out of the 200 word sets, 80 word sets had one of the four target words and 120 word sets did not have any of the target words. On the word sets in which target word was present, presence of a target word in any one of the four cells was equi-

probable. Once the list of 200 word sets was created, each of the word in a word set was then placed in the center of a visual Gaussian noise background. Each of these pictures was saved in the bmp file format. The experiment control program "E-Prime 2.0 (Psychological Software Tools, 2007)" was used to present these visual stimuli along with performing other aspects of experimental control such as reaction time measurement and randomization. Visual stimuli were presented with an HP flat panel, LCD display with a resolution of 1280 X 1024 and a refresh rate of 60 Hz.

Training

Prior to participating in the perceptual experiment, the participants received two one-hour training sessions on visual word recognition task alone and two one-hour training on performing visual word recognition and adaptive pursuit rotor tasks simultaneously. The training session for visual word recognition consisted of two sessions of one hour each. Each session consisted of 5 blocks of visual word recognition trials with a short break between trials. Each block consisted of 200 trials of visual word recognition. The number of target words was set to four words. All the participants practiced with the same target words. The target words were "PATH", "HEAP", "SHED", and "BUSH". The participants were asked to respond as rapidly as possible to the visual stimuli without sacrificing the accuracy. Also, the participants were asked to use their non-dominant hand to make the response.

Participants were required to memorize the four target words before the start of every training session. Once the participants felt they had memorized the target words (unlimited study time was permitted), they pushed a button to initiate the trial. Thereafter, a fixation dot was presented for 500 ms followed by the test frame display. The test frame was displayed for a maximum of 2 seconds or until the participant responded. The test frame display consisted of four words in a visual Gaussian noise background arranged in a 2 X 2 matrix. All the words on the test frame were four letter words.

To illustrate, the participant was presented with the target words *PATH, HEAP, SHED,* and *BUSH* and were asked to memorize them. Once the participant memorized the target words, a button was pushed to initiate the trial. After the presentation of the fixation dot for 500 ms, the participant was presented with the words *TEST, FORK, PEST* and *BUSH* in a visual Gaussian noise background. The participant was instructed to press the button labeled "Present" if one of the four target words were present and "Absent" if none of the four target words were present.

The participants were provided with two feedback cues: (a) on correct responses, the participants saw a picture with words "correct response" and the response button which they pressed. (b) On incorrect responses, the participants saw a picture with words "incorrect response", the response button which they pressed and the correct response button. (c) The

participants were also provided with an indication of their cumulative accuracy during training. The feedback procedure was used to motivate performance over the thousands of trials during the training sessions.

During the training for visual word identification and adaptive pursuit rotor, the participants performed word identification task using their non-dominant hand and adaptive pursuit rotor task using their dominant hand. This training lasted for two one hour sessions. Each session consisted of 10 blocks of trails with a short break between trials. Each trial consisted of participants tracking the moving target for 20 revolutions on the adaptive pursuit rotor task.

Experimental set-up for visual word recognition task

A Dell computer running the Microsoft Windows XP operating system was used to display the visual stimuli. Random presentation of visual stimuli was controlled by E-Prime 2.0 (Psychology Software Tools, Inc., PA, USA). The stimuli were displayed on a 15-inch flat screen monitor set to a resolution of 1024 X 768 pixels @60 Hz. A response box with a serial adaptor was connected to the computer and used to control the E-Prime based visual word recognition task. The response box had two buttons with one labeled as "Present" and the other as "Absent".

Procedure for visual word recognition task

Each participant performed the visual word recognition task along with the adaptive pursuit rotor task and auditory word repetition task simultaneously. The visual stimulus appeared in the centre of the noise background for 500 milliseconds. Participants were instructed to press the button labeled "Present" if one of the four target words were present and "Absent" if none of the target words were present. Before starting the visual word recognition task, the participants were provided with oral and written instruction and were familiarized with the stimuli. Figure 23 shows the sequence of events in each trial.

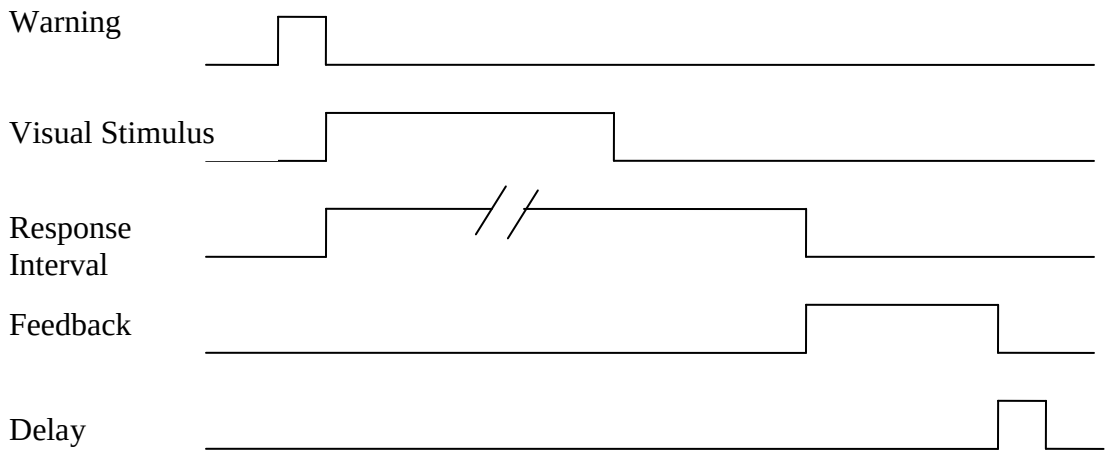

Figure 23. State diagram of visual word recognition task

Procedure

When participants arrived for the experiment, they completed a questionnaire accessing their hearing status, cell phone usage and mouse usage, demographic information, and their driving habits with cell phone. Participants were then familiarized with the adaptive pursuit rotor and visual word recognition tasks. The participants then heard the experimenter explaining the experimental procedure and watched the experimenter doing the adaptive rotor task. The participants then discussed any doubts regarding the experimental procedure with the experimenter.

During the dual-task phase of the experiment, SPIN sentences were presented to the participants via circumaural headphones (Sennheiser HD 520) at approximately 68 dB SPL. The participants were required to repeat the last word of the sentence which they just heard. While doing the listening task, the participants had to concurrently perform the adaptive pursuit rotor task using their dominant hand. The non-dominant hand was used to press the "Present" or "Absent" buttons of the visual word recognition task. Figure 24 shows the entire process of synchronization between the experimental tasks. Figure 25 shows the experimental conditions for the consistent mapping and varied mapping tasks. Figure 26 shows the state diagram for the perceptual phase of the experiment. Figure 27 shows a sample stimulus for the visual word recognition part of experiment 3.

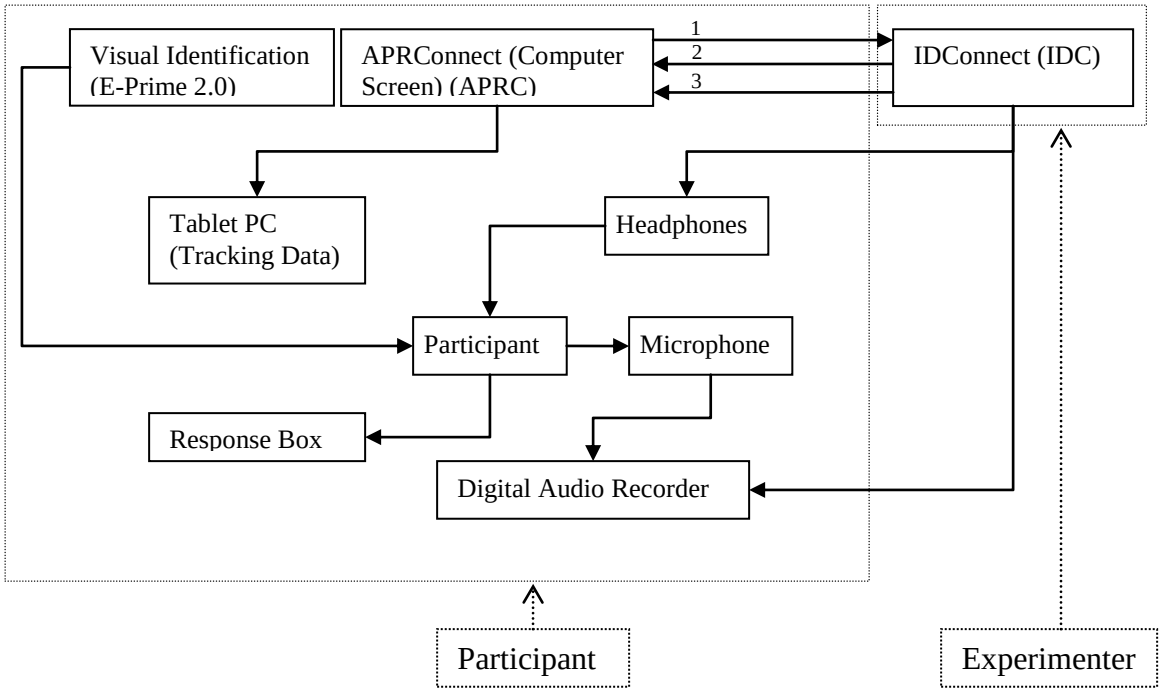

Figure 24. Experimental setup for different experimental tasks.

Consistent Mapping				Varied Mapping			
Session #	Visual ID	APR	Auditory	Session #	Visual ID	APR	Auditory
1	1	--	--	1	1	--	--
2	1	--	--	2	1	--	--
3	1	✓	--	3	1	✓	--
4	1	✓	--	4	1	✓	--
5	1	✓	✓	5	2	✓	✓
6	2	✓	✓	6	1	✓	✓

Visual ID 1 → Target words same during all blocks. Target words are *PATH*, *HEAP*, *SHED*, and *BUSH*.

Visual ID 2 → Target words vary from block to block.

Figure 25. Experimental conditions for Consistent Mapping and Varied Mapping Tasks

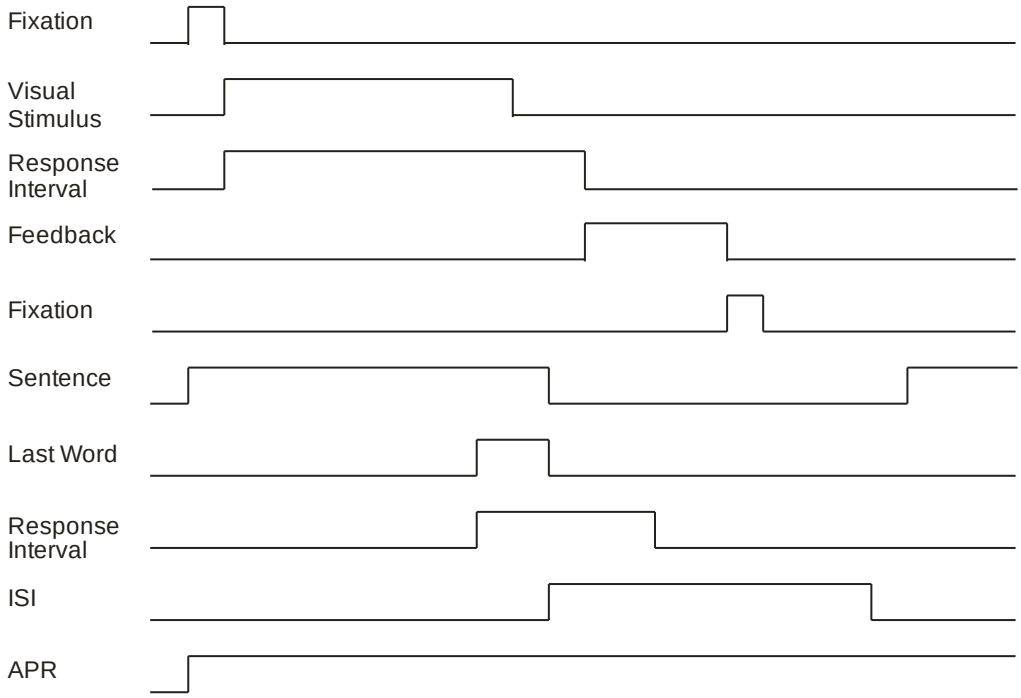

Figure 26. State diagram of perceptual phase

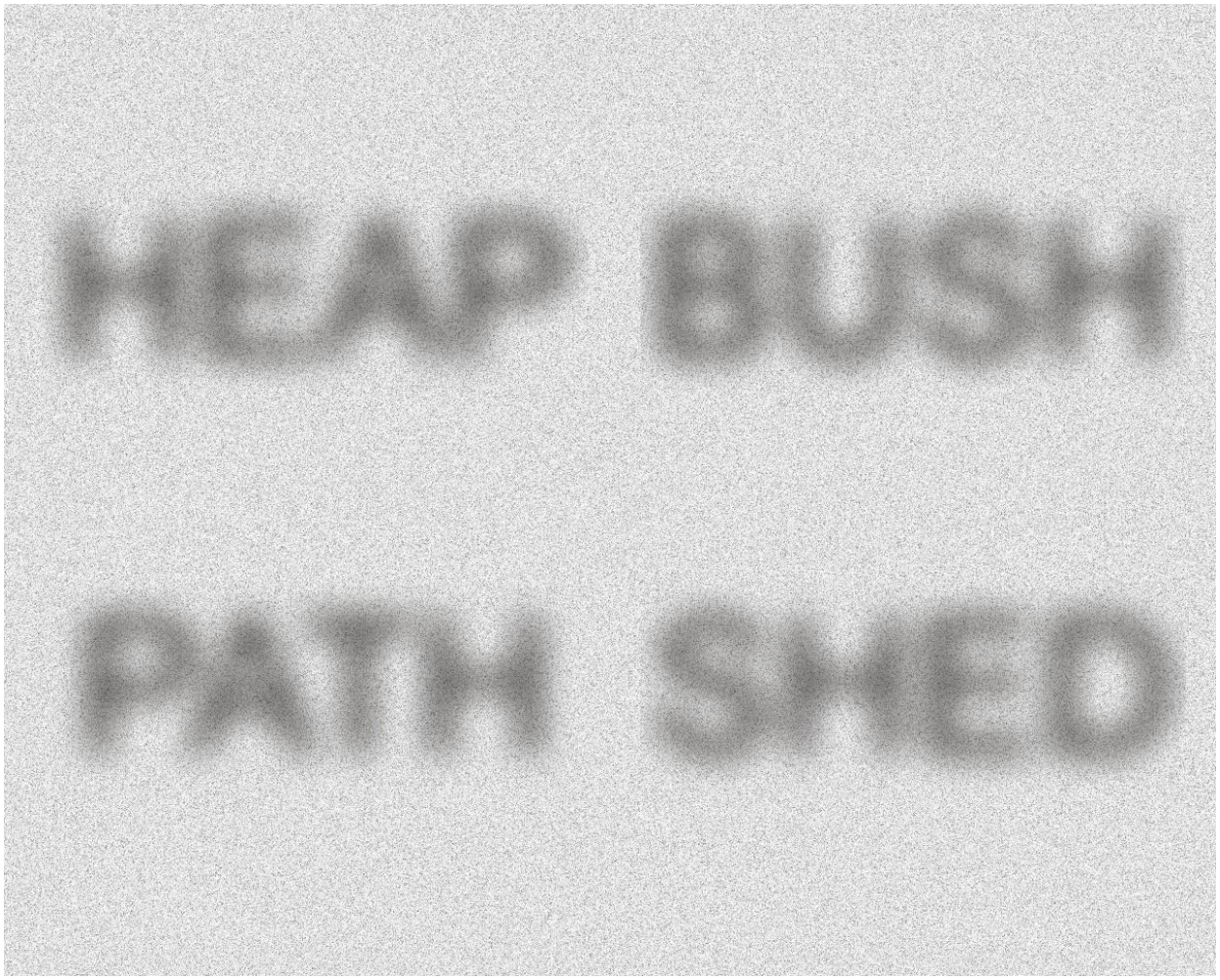

Figure 27. Visual word recognition stimulus

Dependent Measures

During the experiment, APRConnect collected the tracking data which consisted of performance of participants on the tracking task. Also, a digital recorder (Marantz 660 – Solid state recorder) was used to record the sentences and the participant's vocal responses in two different channels. The left channel in the recording recorded the auditory stimulus presented to the participants. The right channel in the recording recorded the verbal

responses of the participant. E-Prime collected data for the consistent mapping or varied mapping visual word recognition task. From this data, the following five performance variables were extracted:

Reaction Time – Word Repetition: Reaction time for word repetition was measured as the time elapsed between the end of the sentence presented to the participant and to the moment when the participant starts answering. Shorter reaction time represented better performance.

Percent Correct – Word Repetition: Percent correct for word repetition represented the proportion of the final words of the sentences correctly identified by the participant. Identifying the last word to be a plural when it was actually singular and vice versa were considered to be incorrect responses. Higher percent correct represented better performance.

Average Speed: Average speed was the average speed in rotations per minute at which the participant did the tracking task using moving average parameters. Faster speed represented better performance.

Reaction time(for hits) – Visual word recognition: Reaction time (hits) for visual word recognition was measured as the time elapsed between the presentation of a visual stimulus containing one of the four target words to the participant pressing the "present" button on the response box. Shorter time represented better performance.

Reaction time(for correct rejection) – Visual word recognition: Reaction time (correct rejection) for visual word recognition was measured as the time

elapsed between the presentation of a visual stimulus that did not contain any of the four target words to the participant pressing the "absent" button on the response box. Shorter time represented better performance.

Percent Correct – Visual word recognition: Percent correct for visual word recognition was the percentage of the target words correctly identified by the participant during the individual experimental conditions. Higher percent represented better performance.

Results

Comprehensive Results.

The first research question investigated whether the two speech sources used, cell phone speech and natural speech, was processed differentially during consistent mapping and varied mapping tasks. To answer this question, the word accuracy during auditory word repetition task was compared for cell phone speech and natural speech during CM and VM tasks.

A 4-way within groups ANOVA was used to examine the main effects and interactions of mapping (consistent and varied), speech (cell phone and natural), level (SN1, SN2, and SN3), and context (predictable and unpredictable) as they relate to word accuracy in the auditory word repetition task.

There was a significant 4-way interaction as they relate to word accuracy identified ($F (2, 30) = 11.1$, $p < 0.001$, $M_{se} = 16.489$). Visual Inspection

of the cell means (using LSD_{mmd} = 2.932) revealed that, when the visual identification task was consistently mapped and as the quality of the presented cell phone signal improved, the amount of last words correctly identified by the participants improved. This was true for both predictable and unpredictable sentences. Also, the participants performed better on predictable sentences as compared to unpredictable sentences as the signal quality improved. However, when the mapping was consistent and the auditory signal presented was natural speech, for predictable sentences, there was no significant difference in the word accuracy between -4 dB SNR and -2 dB SNR. Also, there was a significant difference in the word accuracy between -2 dB SNR and 0 dB SNR. For unpredictable sentences, as the signal quality improved, the amount of last words correctly identified by the participants improved. When the visual identification task was variably mapped, as the signal quality improved, the amount of last words correctly identified by the participants improved. This was true for both cell phone speech and natural speech. Also, the participants performed better for predictable sentences as compared to unpredictable sentences. Table 16 shows the mean word accuracy for predictable and unpredictable sentences at different levels of SNR for cell phone and natural speech during CM and VM tasks. Figure 28 shows mapping by speech by level by context interaction for mean word accuracy during auditory word repetition task. The error bars denote 95% confidence intervals.

Table 16. Means for word accuracy (%) for predictable and unpredictable sentences at levels SN1 SN2, and SN3 for cell phone and natural speech during CM and VM tasks

Consistent Mapping						
	Cell Phone Speech			Natural Speech		
	4 dB	6 dB	8 dB	-4 dB	-2 dB	0 dB
Predictable	75.313	81.563	94.688	83.75	85	96.875
Unpredictable	35.938	40.938	48.75	46.875	55.625	72.813

Varied Mapping						
	Cell Phone Speech			Natural Speech		
	4 dB	6 dB	8 dB	-4 dB	-2 dB	0 dB
Predictable	71.25	76.563	84.688	77.813	81.875	91.875
Unpredictable	31.563	39.375	46.25	38.75	49.063	55

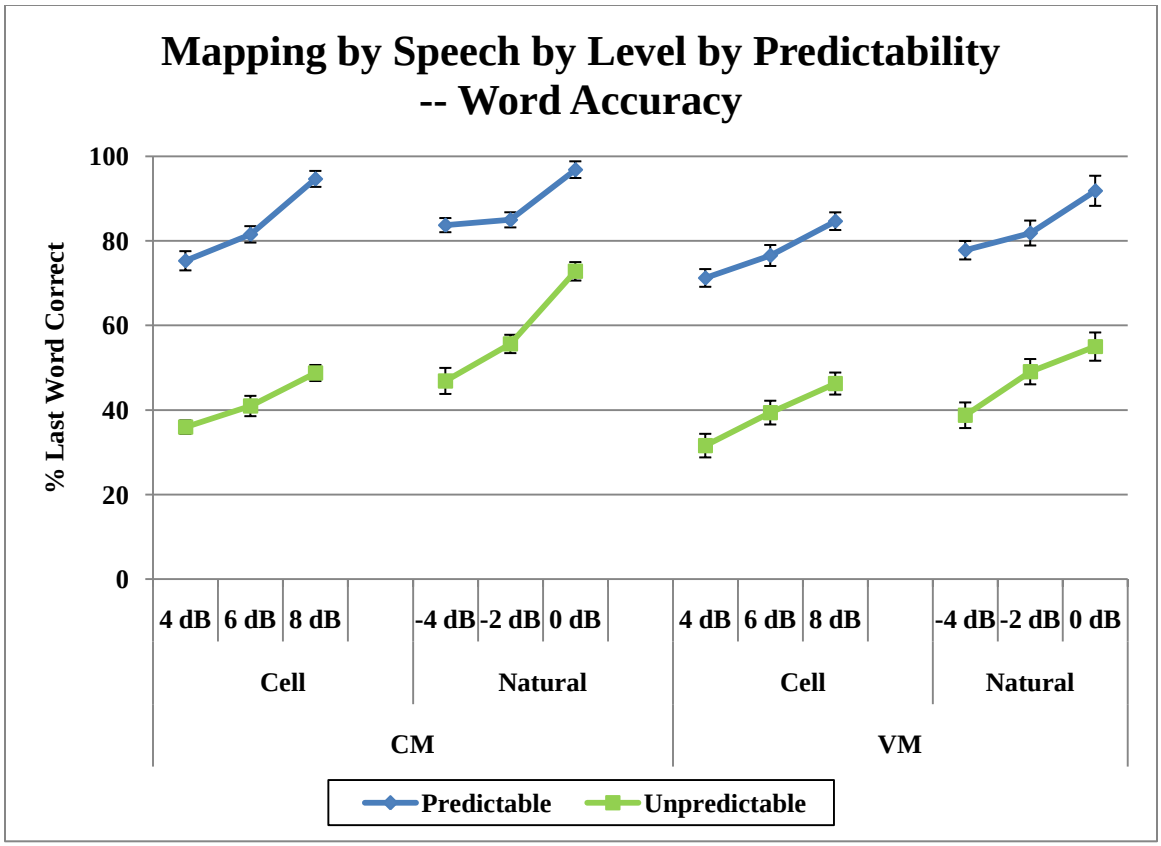

Figure 28. Effect of CM and VM task on word accuracy for cell phone and natural speech at different SNRs

The next research question investigated on the necessity of different attentional mechanisms for cell phone speech and natural speech perception. To answer this question, the reaction time for word repetition was compared for cell phone speech and natural speech during CM and VM tasks.

There was a significant 2-way interaction between mapping and speech as they relate to mean reaction time for word repetition during the auditory word repetition task ($F(1, 15) = 130.3$, $p < 0.001$). Visual Inspection of the cell means revealed that, as expected, the participants had a higher

mean reaction time for auditory word repetition when the task was variably mapped as compared to when the task was consistently mapped. Also, the mean reaction time for auditory word repetition was higher when cell phone speech was presented as compared to natural speech. This pattern of results was obtained for both consistent mapping and varied mapping visual word identification task. The difference in reaction time between consistent mapping and varied mapping task was higher for cell phone speech as compared to natural speech. Table 17 shows the mean reaction time for auditory word repetition during CM and VM tasks when cell phone speech and natural speech was presented to the participants. Figure 29 shows mapping by speech interaction for the mean reaction time for auditory word repetition task. The error bars denote 95% confidence intervals.

Table 17. Means for the reaction time for auditory word repetition (ms) during CM and VM tasks for cell phone speech and natural speech

Consistent Mapping		Varied Mapping	
Cell Phone Speech	Natural Speech	Cell Phone Speech	Natural Speech
552	489	765	552

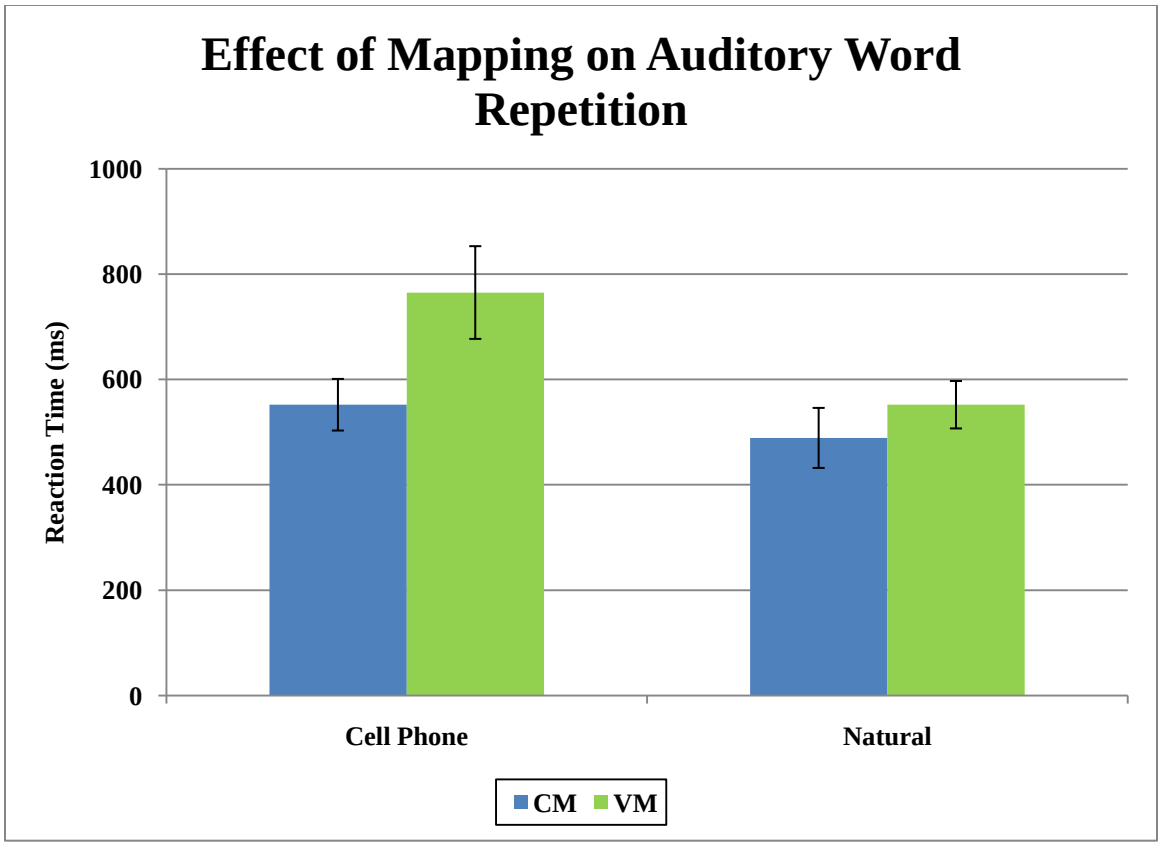

Figure 29. Effect of CM and VM task on mean reaction time for auditory word repetition for cell phone and natural speech

The next research question investigated the effect of consistent mapping and varied mapping tasks on auditory word repetition performance and simultaneous visual-motor performance for cell phone speech and natural speech. To answer this research question, word accuracy and rotation speed was compared for cell phone speech and natural speech during CM and VM tasks.

There was a significant 2-way interaction between mapping (consistent mapping and varied mapping) and speech (cell phone speech and natural

speech) as they relate to average rotation speed ($F(1, 15) = 41.9$, $p < 0.001$, $M_{se} = 0.086$). Visual Inspection of the cell means (using $LSD_{mmd} = 0.09$) revealed that, when cell phone speech was presented to the participants, they had a higher speed of rotation when the visual identification task was consistently mapped as compared to when the visual identification task was variable mapped. However, this relationship reversed when natural speech was presented to the participants. Also, the magnitude of change in average speeds between cell phone and natural speech was higher for consistent mapping as compared to varied mapping. Since the target words for visual identification during consistent mapping task was the same as that of practice phase, the participants spent more resources on the simultaneous visual-motor task as compared to visual word identification task during CM task as compared to VM task. Table 18 shows the average speed of rotation for cell phone and natural speech during CM and VM tasks. Figure 30 shows the mapping by speech interaction. The error bars denote 95% confidence intervals.

Table 18. Means for the average rotation speed (rotations per minute) in visual-motor task during CM and VM tasks for cell phone speech and natural speech

Consistent Mapping		Varied Mapping	
Cell Phone Speech	Natural Speech	Cell Phone Speech	Natural Speech
7.06	6.33	4.94	5.53

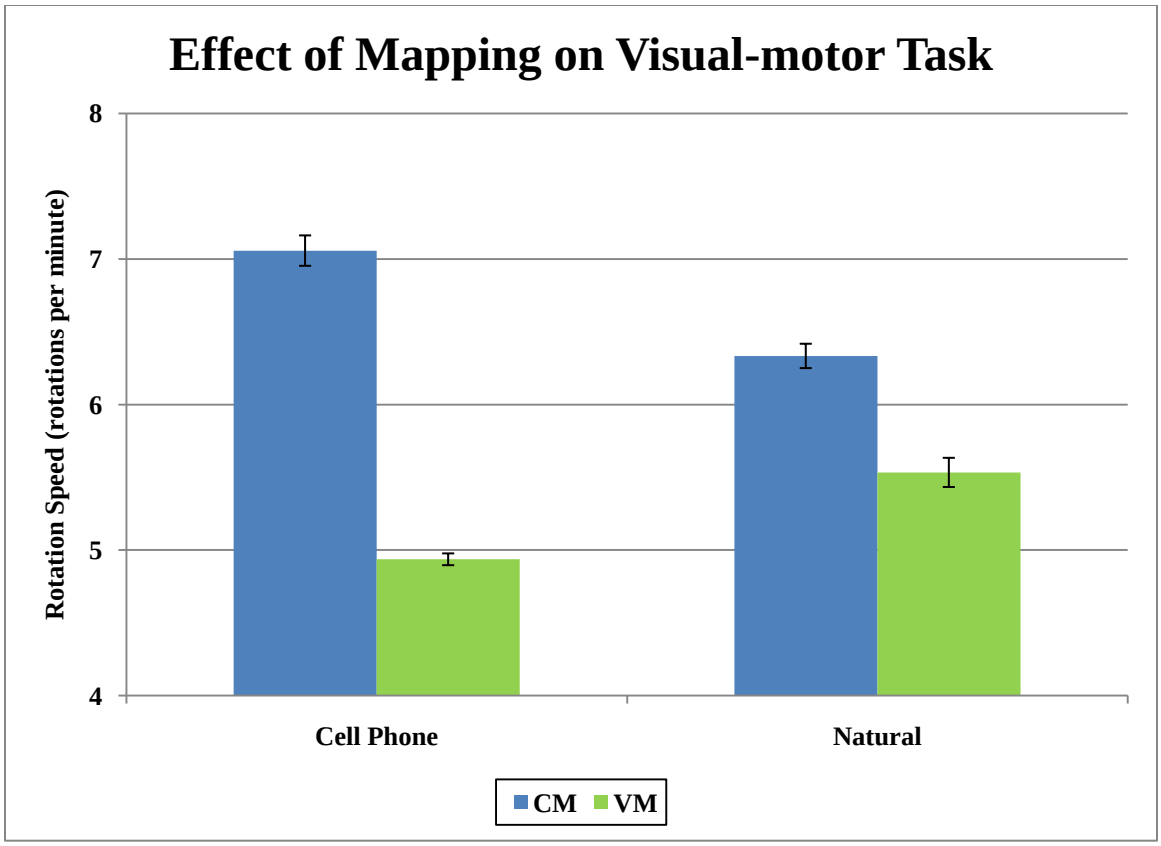

Figure 30. Effect of CM and VM task on simultaneous visual-motor task performance for cell phone and natural speech

There was a significant 2-way interaction between mapping (consistent mapping and varied mapping) and speech (cell phone speech and natural speech) as they relate to word accuracy during auditory word repetition task ($F(1, 15) = 9.6$, $p = 0.007$, $M_{se} = 25.308$). Visual Inspection of the cell means (using $LSD_{mmd} = 1.547$) revealed that the average last word correct was always higher when the visual identification task was consistently mapped as compared to when the visual identification task was variable mapped. However, the difference in average last word correct between consistent and varied mapped tasks was higher for natural speech as compared to cell phone

speech. Table 19 shows the word accuracy during auditory word repetition for cell phone and natural speech during CM and VM tasks. Figure 31 shows the mapping by speech interaction. The error bars denote 95% confidence intervals.

Table 19. Means for the word accuracy (%) during auditory word repetition during CM and VM tasks for cell phone speech and natural speech

Consistent Mapping		Varied Mapping	
Cell Phone Speech	Natural Speech	Cell Phone Speech	Natural Speech
62.86	73.79	58.28	65.73

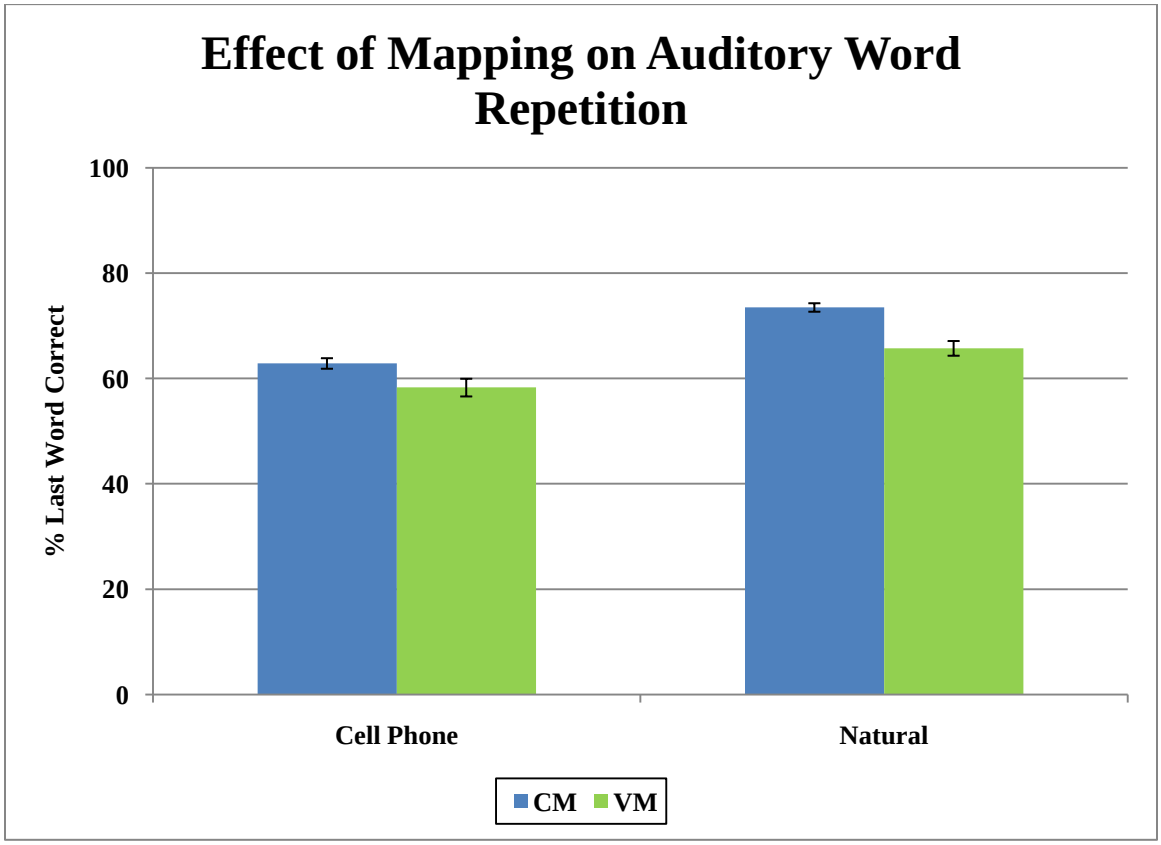

Figure 31. Effect of CM and VM tasks on auditory word repetition task for cell phone and natural speech

The next research question investigated the effect on simultaneous visual-motor task performance due to the addition of simultaneous visual word identification task. To answer this research question, average speed on the simultaneous visual-motor task was compared for cell phone speech and natural speech during CM and VM tasks.

A 4-way within groups ANOVA was used to examine the main effects and interactions of mapping (consistent and varied), speech (cell phone and natural), level (4 dB, 6 dB, and 8 dB for cell phone speech; -4 dB, -2 dB, and 0 dB for natural speech), and context (predictable and unpredictable) as they relate to the average rotation speed in the adaptive pursuit rotor task.

There was a significant 4-way interaction as they relate to adaptive tracking performance ($F(2, 30) = 25.5$, $p < 0.001$, $M_{se} = 0.088$). Visual Inspection of the cell means (using $LSD_{mmd} = 0.212$) revealed that, when the visual identification task was consistently mapped and the auditory signal presented was cell phone speech, at 4 dB and 6 dB SNRs, the participants had a higher average speed of rotation for unpredictable sentences as compared to predictable sentences. However, at 8 dB SNR, participants had a higher average speed of rotation for predictable sentences as compared to unpredictable sentences. However, when the mapping was consistent and the auditory signal presented was natural speech, there was no significant difference in speed of rotation between predictable and unpredictable sentences for -4 dB and -2 dB SNR. When the signal presented was at 0 dB

SNR, the participants had a higher rotation speed for unpredictable sentences compared to predictable sentences. When the visual identification task was variably mapped, there was no significant difference in speed of rotation for predictable or unpredictable sentences. This pattern of results was obtained when the auditory stimuli presented were cell phone speech and when they were natural speech. Table 20 shows average rotation speed for predictable and unpredictable sentences at different levels of SNR for cell phone and natural speech during CM and VM tasks. Figure 32 shows mapping by speech by level by context interaction for average rotation speed during adaptive pursuit rotor task. The error bars denote 95% confidence intervals.

Table 20. Means for average rotation speed (rotations per minute) for predictable and unpredictable sentences at different levels for cell phone and natural speech during CM and VM tasks

Consistent Mapping						
	Cell Phone Speech			Natural Speech		
	4 dB	6 dB	8 dB	-4 dB	-2 dB	0 dB
Predictable	6.55	9.92	7.39	6.59	6.38	6.01
Unpredictable	7.38	7.15	6.95	6.4	6.35	6.28

Varied Mapping						
	Cell Phone Speech			Natural Speech		
	4 dB	6 dB	8 dB	-4 dB	-2 dB	0 dB
Predictable	5.17	5.05	4.79	5.43	5.54	5.61
Unpredictable	5.00	4.93	4.68	5.62	5.53	5.47

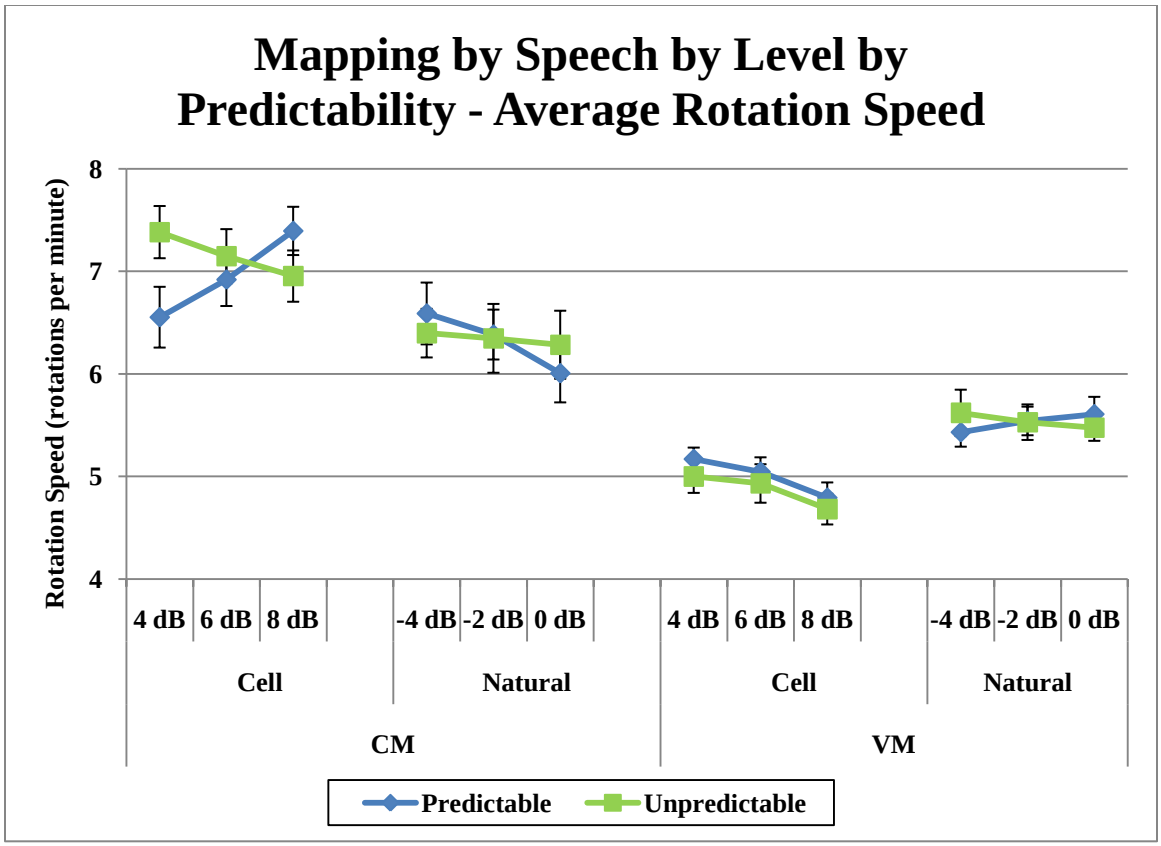

Figure 32. Effect of CM and VM task on simultaneous visual-motor performance for cell phone and natural speech at different SNRs

The next research question investigated the effect on visual word identification task performance during cell phone speech and natural speech perception. To answer this question, the reaction time for hits and reaction time for correct rejection for the visual word identification task was compared for cell phone speech and natural speech during CM and VM tasks.

There was a significant 2-way interaction between mapping (consistent mapping and varied mapping) and speech (cell phone speech and natural speech) as they relate to mean reaction time for hits during visual word

identification task ($F(1, 15) = 986.7$, $p < 0.001$, $M_{se} = 6824.264$). Visual Inspection of the cell means (using $LSD_{mmd} = 25.409$) revealed that, when cell phone speech was presented to the participants, they had a higher reaction time for hits when the visual identification task was consistently mapped as compared to when the visual identification task was variable mapped. this might be due to the fact that the target words during CM task was well learnt and the participants probably spent less resources on the visual word identification task and more resources to the auditory word repetition task and hence had a higher reaction time for hits. When the task was VM, the participants spent more resources on the visual word identification task and hence had a lower reaction time for hits. However, when, natural speech was presented to the participants, they had a lower reaction time for hits when the visual identification task was consistently mapped as compared to when the visual identification task was variably mapped. Also, the magnitude of change in reaction time was higher when the identification task was variable mapped as compared to consistently mapped task. Table 21 shows the mean reaction time for hits during visual word identification task for cell phone and natural speech during CM and VM tasks. Figure 33 shows the mapping by speech interaction. The error bars denote 95% confidence intervals.

Table 21. Means for the average reaction time for hits (ms) during CM and VM tasks for cell phone speech and natural speech

Consistent Mapping		Varied Mapping	
Cell Phone Speech	Natural Speech	Cell Phone Speech	Natural Speech
1540	1373	1236	1599

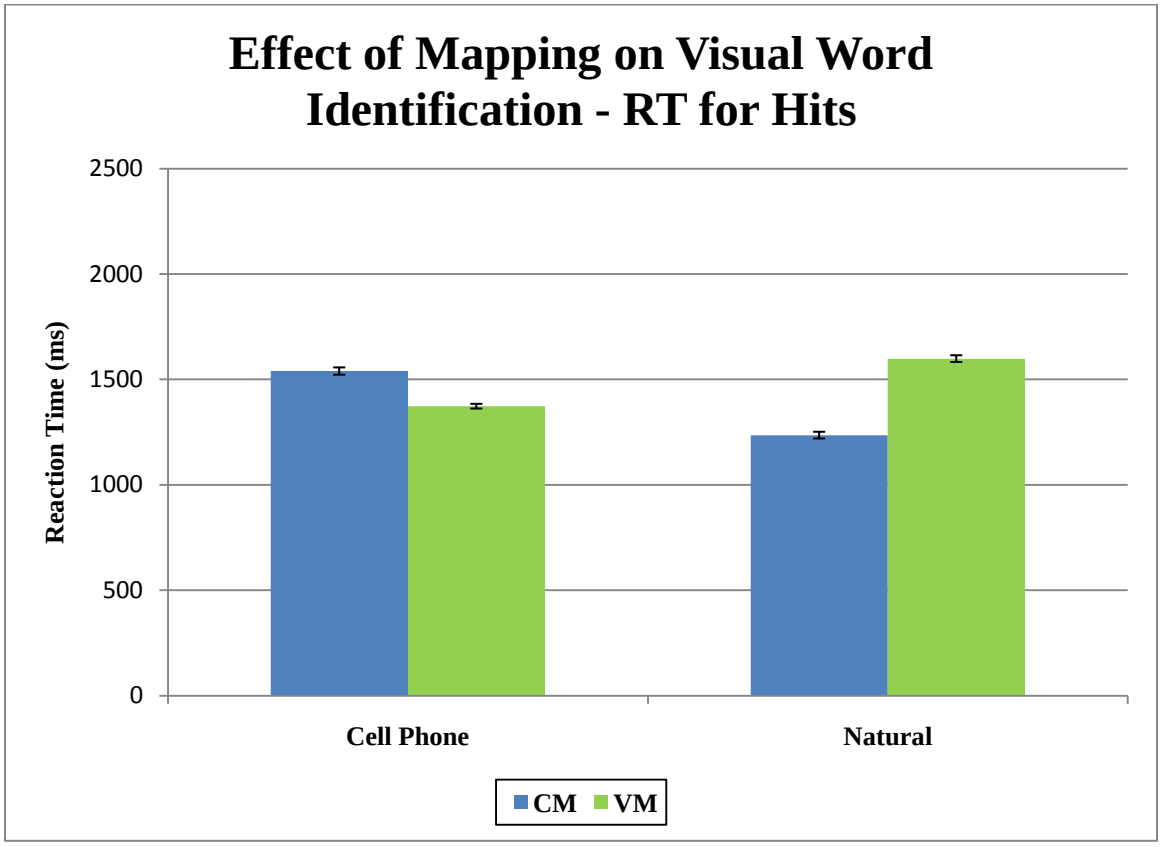

Figure 33. Effect of CM and VM tasks on visual word identification task (hits) for cell phone and natural speech

There was a significant 2-way interaction between mapping (consistent mapping and varied mapping) and speech (cell phone speech and natural speech) as they relate to mean reaction time for correct rejection during

visual identification task ($F(1, 15) = 23.5$, $p < 0.001$, $M_{se} = 20409.530$). Visual Inspection of the cell means (using $LSD_{mmd} = 43.902$) revealed that the participants had a higher reaction time for correct rejection when the visual identification task was variably mapped as compared to when the visual identification task was consistently mapped. This was true when both cell phone speech and natural speech was presented to the participants. When the visual identification task was consistently mapped, the participants had a higher reaction time for correct rejection for cell phone speech as compared to natural speech. However, when the visual identification task was variably mapped, the participants had a higher reaction time for correct rejection for natural speech as compared to cell phone speech. Table 22 shows the mean reaction time for correct rejection during visual word identification task for cell phone and natural speech during CM and VM tasks. Figure 34 shows the mapping by speech interaction. The error bars denote 95% confidence intervals.

Table 22. Means for the average reaction time for hits (ms) during CM and VM tasks for cell phone speech and natural speech

Consistent Mapping		Varied Mapping	
Cell Phone Speech	Natural Speech	Cell Phone Speech	Natural Speech
1634	1543	2013	2063

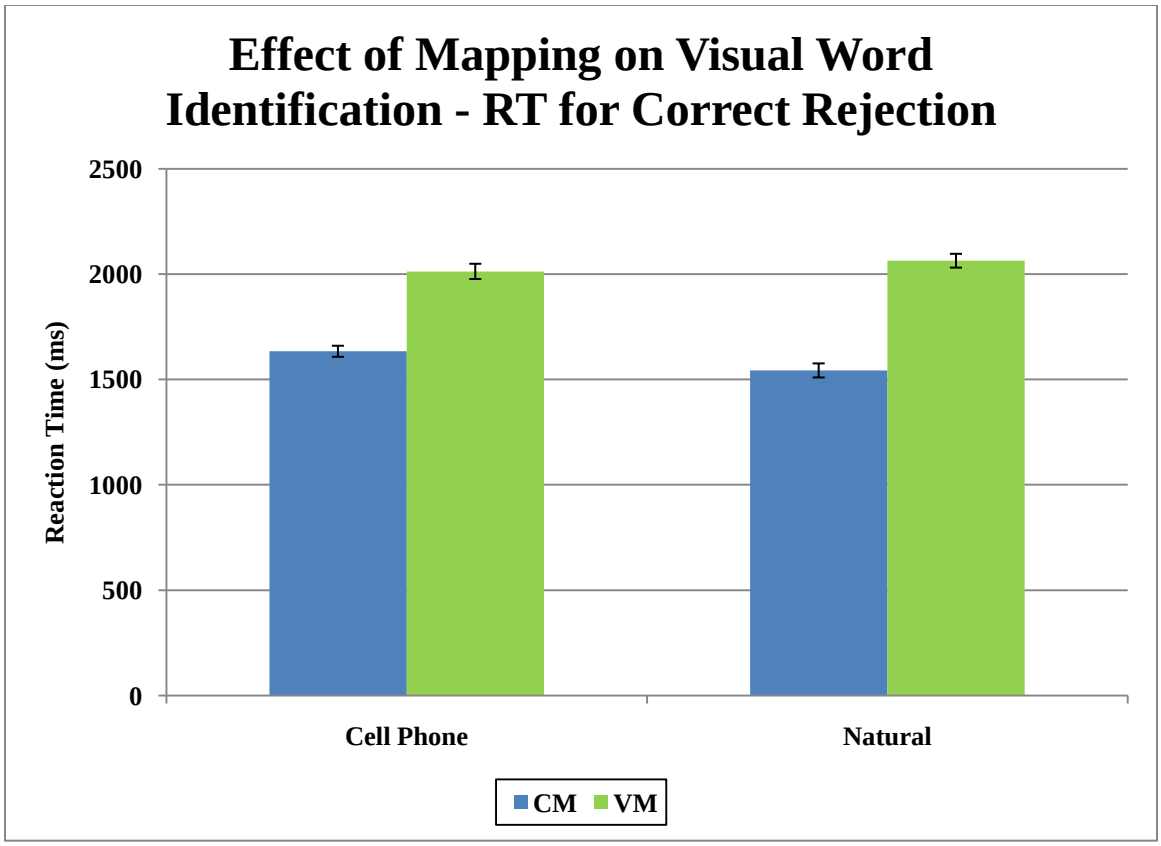

Figure 34. Effect of CM and VM tasks on visual word identification task (correct rejection) for cell phone and natural speech

Other Results

There was a significant 3-way interaction between speech (cell phone speech and natural Speech), context (predictable and unpredictable), and level (4 dB, 6 dB, and 8 dB for cell phone speech; -4 dB, 2 dB, and 0 dB for natural speech) (F (2, 30) = 8.8, p = 0.001, M_{se} = 0.107). Visual Inspection of the cell means (using LSD_{mmd} = 0.167) revealed that, when the auditory signal was cell phone speech, at 4 dB SNR, participants had a higher speed of rotation for unpredictable sentences as compared to predictable sentences. However, at 6 dB SNR, there was no significant difference in speeds for

predictable and unpredictable sentences. At 8 dB SNR, participants had a higher speed of rotation for predictable sentences as compared to unpredictable sentences. Also, when the auditory signal presented was natural speech, there was no significant difference in the speed of rotation between predictable or unpredictable sentences. Table 23 shows the average speed of rotation for predictable and unpredictable sentences at different levels of SNR for cell phone and natural speech. Figure 35 shows the speech by level by context interaction. The error bars denote 95% confidence intervals.

Table 23. Means for average rotation speed (rotations per minute) for predictable and unpredictable sentences at different SNR levels for cell phone and natural speech

	Cell Phone Speech			Natural Speech		
	4 dB	6 dB	8 dB	-4 dB	-2 dB	0 dB
Predictable	5.860938	5.982188	6.092813	6.01	5.961875	5.806563
Unpredictable	6.190625	6.039063	5.817188	6.008125	5.9375	5.879063

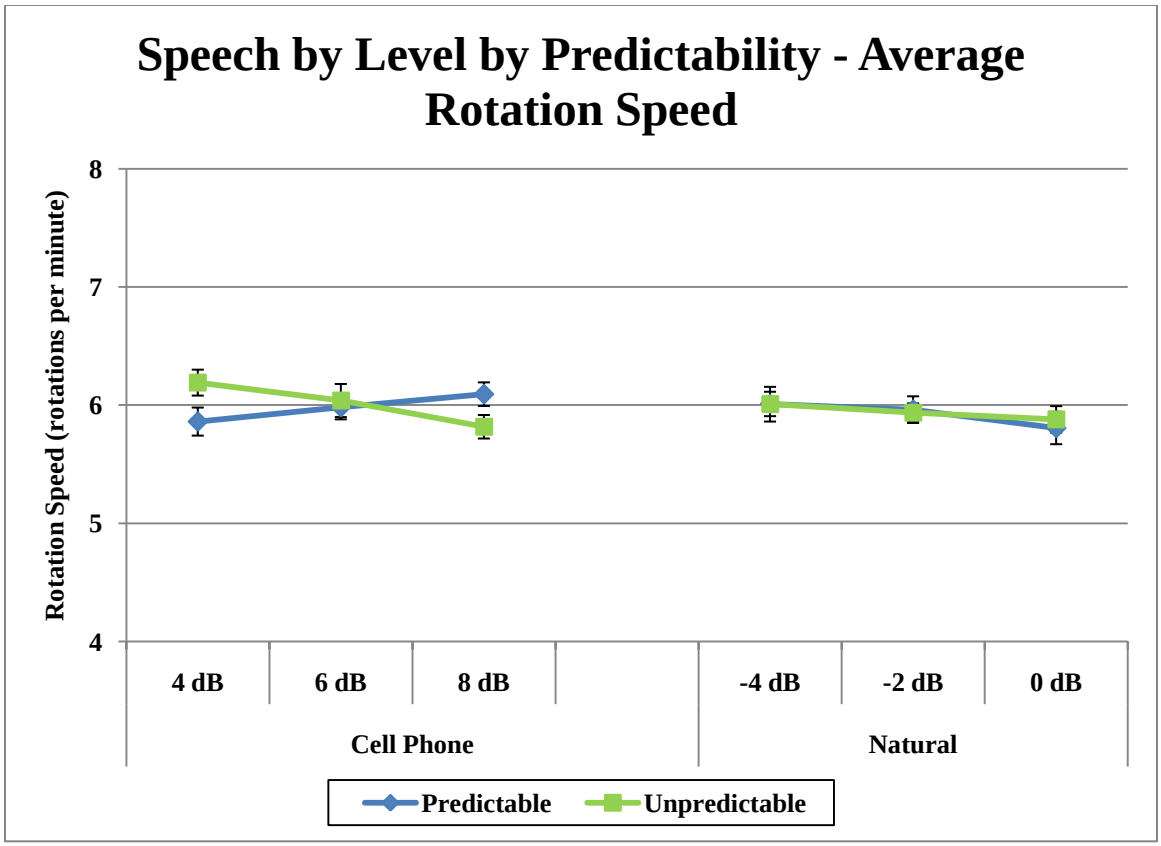

Figure 35. Effect of cell phone speech and natural speech at different SNRs on adaptive tracking task for predictable and unpredictable sentences

There was a significant 2-way interaction between mapping (consistent mapping and varied mapping) and context (predictable and unpredictable) as they relate to average rotation speed ($F(1, 15) = 7.2$, $p = 0.017$, $M_{se} = 0.095$). Visual Inspection of the cell means (using $LSD_{mmd} = 0.094$) revealed that, predictable sentences had a higher speed of rotation than unpredictable sentences for both consistent mapping and varied mapping task. However, the difference between predictable and unpredictable speeds was higher for consistent mapping task as compared to varied mapping task. Also,

predictable sentences had a higher difference in speed between CM and VM conditions as compared to unpredictable sentences. Table 24 shows the average speed of rotation for predictable and unpredictable sentences during CM and VM tasks. Figure 36 shows the mapping by context interaction. The error bars denote 95% confidence intervals.

Table 24. Means for the average rotation speed (rotations per minute) in visual-motor task during CM and VM tasks for predictable and unpredictable sentences

Consistent Mapping		Varied Mapping	
Predictable	Unpredictable	Predictable	Unpredictable
6.6406	6.7515	5.2642	5.2057

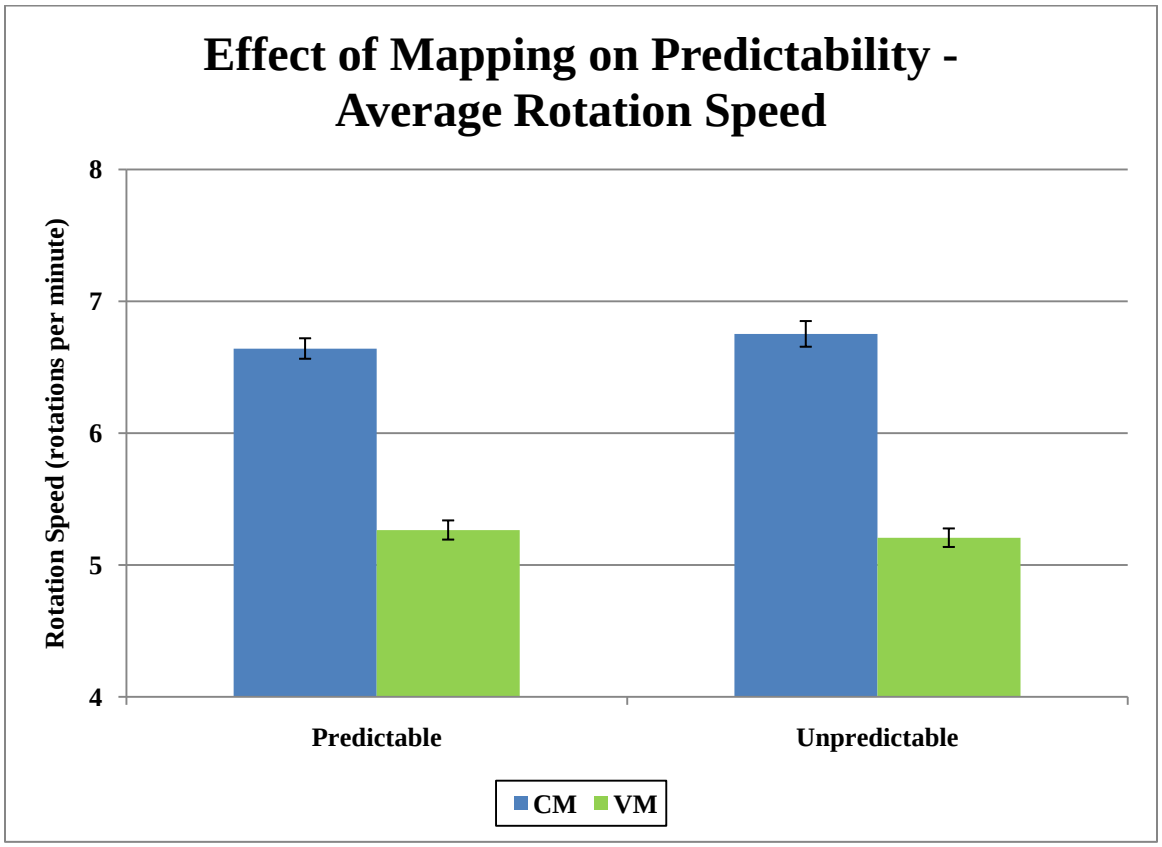

Figure 36. Effect of CM and VM tasks on simultaneous visual-motor task performance for predictable and unpredictable sentences

There was a significant 2-way interaction between mapping (consistent mapping and varied mapping) and predictability (predictable and unpredictable) as they relate to mean reaction time for hits during visual identification task ($F(1, 15) = 4.8$, $p = 0.045$, $M_{se} = 5732.401$). Visual Inspection of the cell means (using $LSD_{mmd} = 23.289$) revealed that the predictable sentences had a lower meant reaction time for hits as compared to unpredictable sentences. However, the difference in the reaction times between predictable and unpredictable sentences was higher when the visual identification task was consistently mapped as compared to when the visual identification task was variably mapped. Table 25 shows the average reaction for hits for predictable and unpredictable sentences during CM and VM tasks. Figure 37 shows the mapping by context interaction. The error bars denote 95% confidence intervals.

Table 25. Means for average reaction time for hits (ms) for predictable and unpredictable sentences during CM and VM tasks

Consistent Mapping		Varied Mapping	
Predictable	Unpredictable	Predictable	Unpredictable
1357	1418	1472	1499

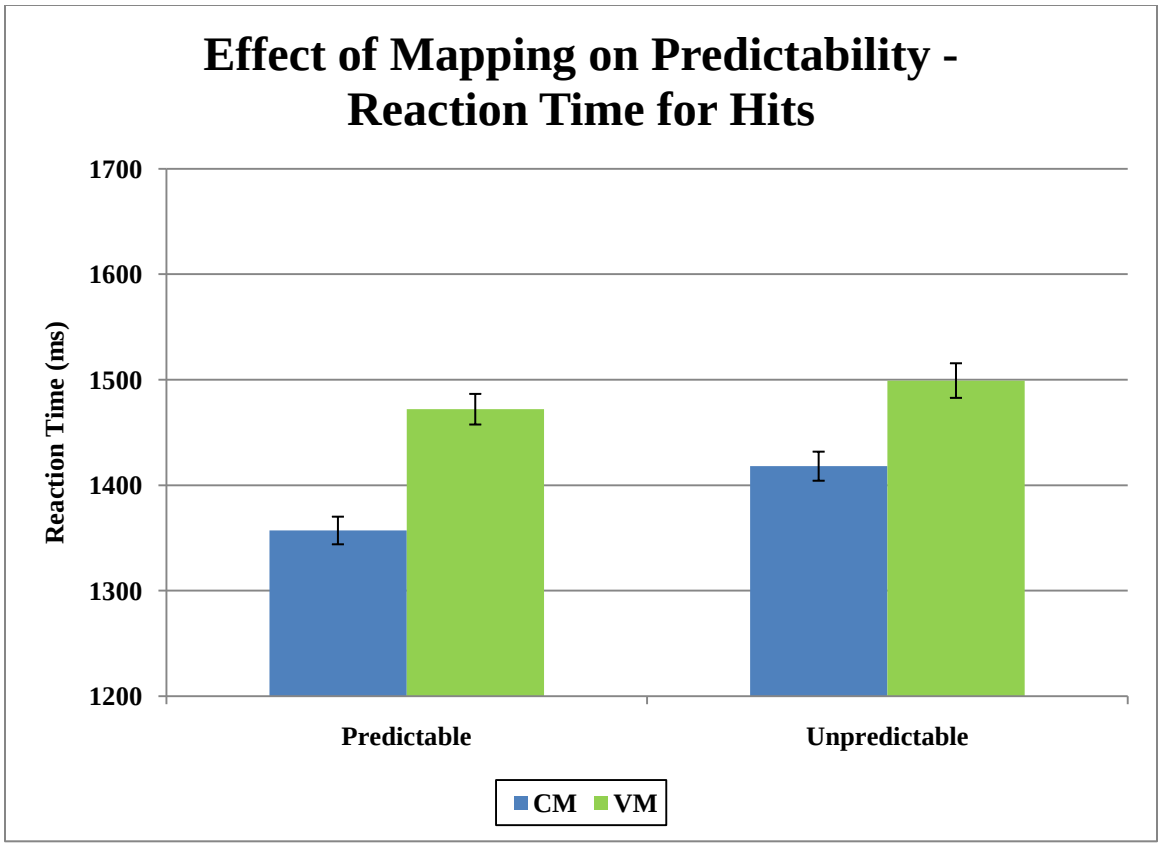

Figure 37. Effect of CM and VM tasks on mean reaction time for hits for predictable and unpredictable sentences

There was a significant 2-way interaction between speech (cell phone speech and natural speech) and context (predictable and unpredictable) as they relate to mean reaction time for hits during visual identification task ($F(1, 15) = 11.9$, $p = 0.003$, $M_{se} = 12303.696$). Visual Inspection of the cell means (using $LSD_{mmd} = 34.118$) revealed that when the auditory stimuli was cell phone speech, predictable sentences had lower mean reaction time for hits as compared to unpredictable sentences. However, when natural speech was presented, there was no significant difference in mean reaction time for hits between predictable and unpredictable sentences. Also, there was no

difference in mean reaction time for hits for predictable sentences between cell phone speech and natural speech. When unpredictable sentences were presented, the participants had a higher reaction time for hits for cell phone speech as compared to natural speech. Table 26 shows the average reaction time for hits for predictable and unpredictable sentences for cell phone and natural speech. Figure 38 shows the speech by context interaction. The error bars denote 95% confidence intervals.

Table 26. Means for average reaction time for hits (ms) for predictable and unpredictable sentences for cell phone speech and natural speech

Cell Phone Speech		Natural Speech	
Predictable	Unpredictable	Predictable	Unpredictable
1414	1498	1415	1419

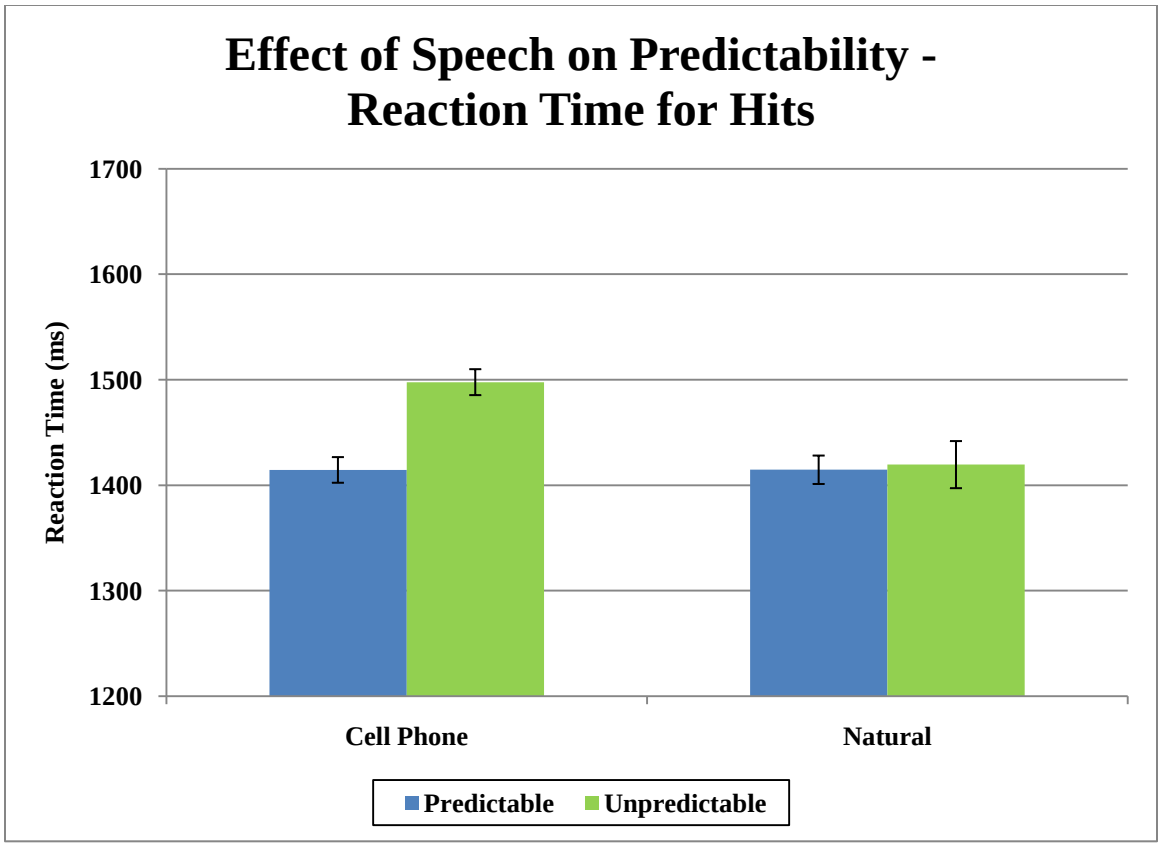

Figure 38. Effect of cell phone speech and natural speech on mean reaction time for hits for predictable and unpredictable sentences

There was a significant 3-way interaction between mapping (consistent mapping and varied mapping), speech (cell phone speech and natural speech), and context (predictable and unpredictable) as they relate to word correctly identified in the auditory word repetition task ($F(1, 15) = 17.8$, $p = 0.001$, $M_{se} = 31.697$). Visual Inspection of the cell means (using $LSD_{mmd} = 2.449$) revealed that the amount of last words correctly identified by the participants was higher for natural speech as compared to cell phone speech. This was true for both consistent mapping and varied mapping visual identification tasks. Also, the participants performed better when the sentences were predictable

as compared to unpredictable sentences. Table 27 shows the word accuracy for predictable and unpredictable sentences for cell phone and natural speech during CM and VM tasks. Figure 39 shows mapping by speech by context interaction. The error bars denote 95% confidence intervals.

Table 27. Means for the word accuracy (%) during auditory word repetition correct for predictable and unpredictable sentences for cell phone and natural speech during CM and VM tasks

	CM		VM	
	Cell Phone	Natural	Cell Phone	Natural
Predictable	83.85	88.54	77.5	83.85
Unpredictable	41.875	58.4375	39.0625	47.60

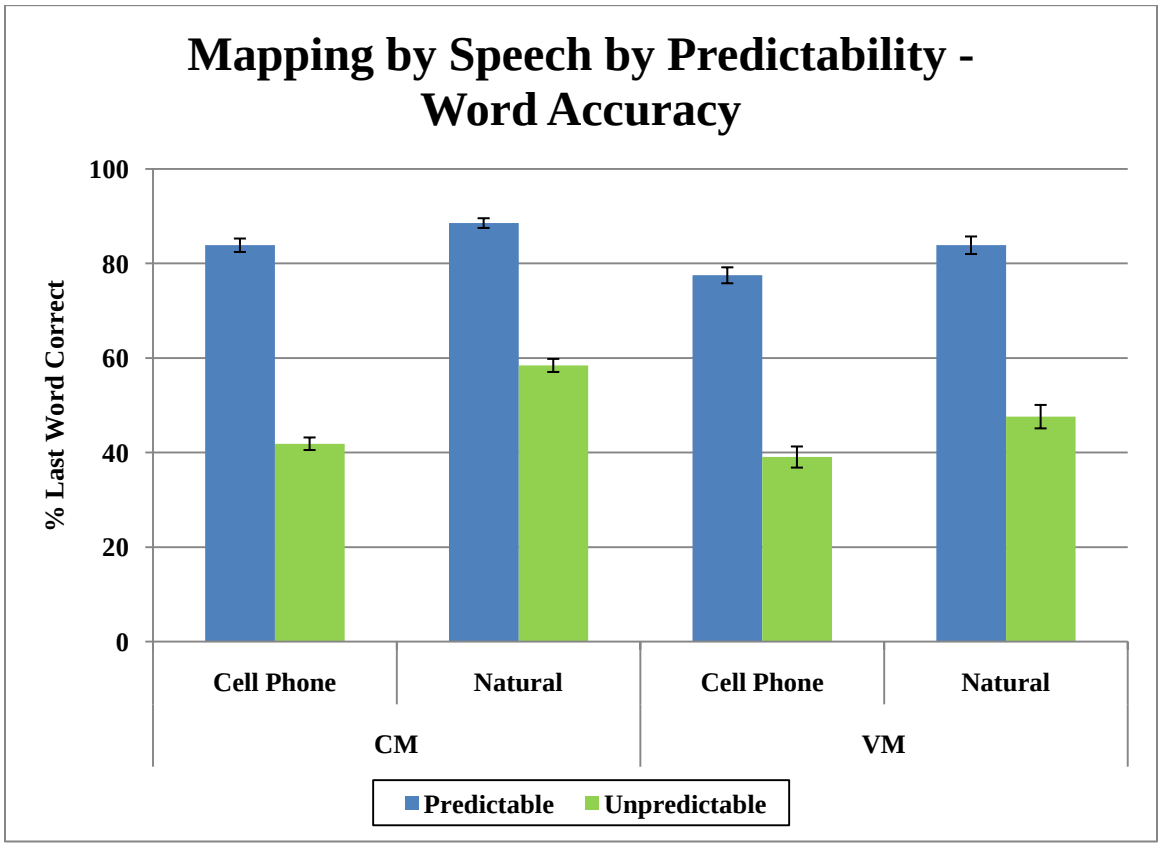

Figure 39. Effect of cell phone speech and natural speech during CM and VM tasks on word accuracy for predictable and unpredictable sentences

Comprehensive Findings.

When APR performance, speech intelligibility, and reaction time for hits in visual word identification task were compared for predictable sentences under consistent mapping condition, it was found that when participants were listening to cell phone speech, increased intelligibility led to an increase in average rotation speed and an increase in reaction time for hits. This showed that at lower intelligibility levels, the participants were paying more attention to the visual identification task as compared to the adaptive pursuit rotor task. However, as the intelligibility of the sentences presented increased, the participants paid more attention to the adaptive pursuit rotor task and less attention to the visual identification task.

When natural speech was presented, increased intelligibility led to a decrease in average rotation speed and a decrease in reaction time for hits. This indicated the fact that as intelligibility of the sentences presented increased, the participants paid less attention to the adaptive pursuit rotor task and more attention to the visual identification task.

Hence, on the whole, when the participants were performing a consistent mapped task and when the speech presented was predictable, listening to cell phone speech had a harmful effect on visual word identification task whereas listening to natural speech had a harmful effect on performance on adaptive pursuit rotor task. Also, listening to cell phone speech had a beneficial effect on the performance on adaptive pursuit rotor

task whereas listening to natural speech had a beneficial effect on visual word identification task. Figure 40 shows the relationship between intelligibility and visual motor performance and reaction time for hits and visual motor performance for predictable sentences during consistent mapping condition

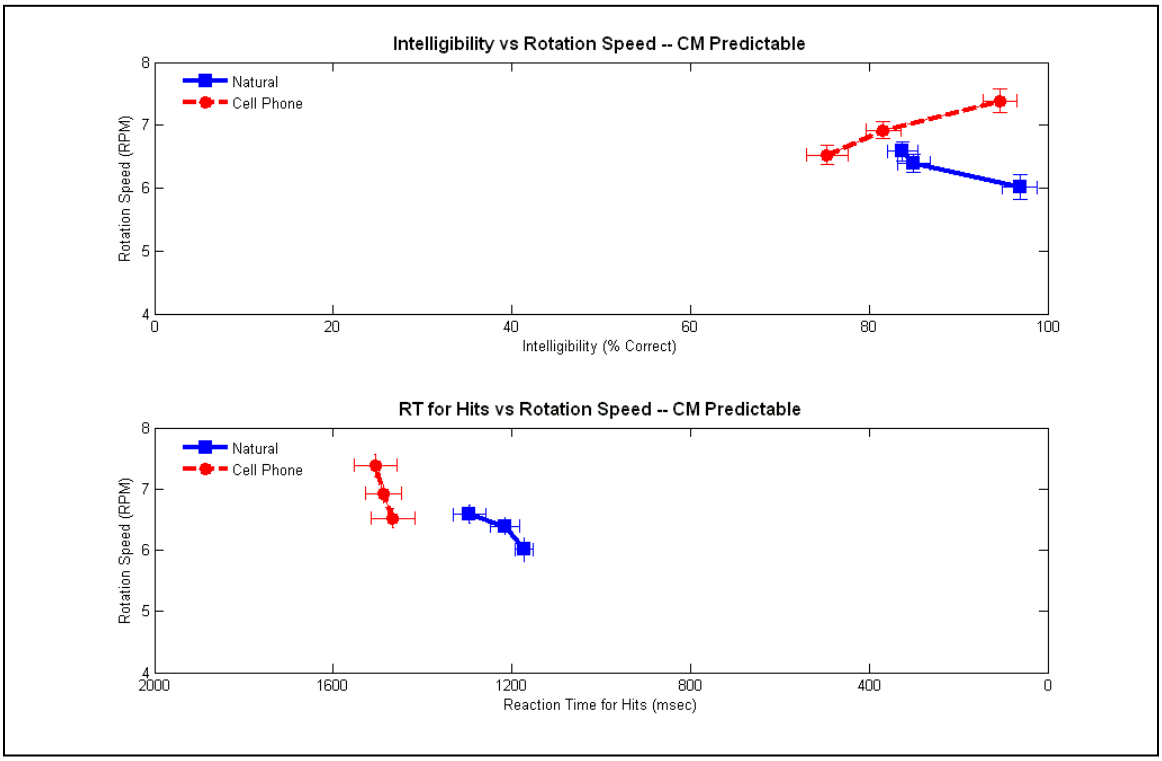

Figure 40. Relationship between intelligibility and visual motor performance and reaction time for hits and visual motor performance for predictable sentences during consistent mapping condition

When APR performance, speech intelligibility, and reaction time for hits in visual word identification task were compared for unpredictable sentences under consistent mapping condition, it was found that when participants were listening to cell phone speech, increased intelligibility led

to a decrease in average rotation speed and a decrease in reaction time for hits. This showed that at lower intelligibility levels, the participants paid more attention to the APR task and less attention to the visual identification task. However, as the intelligibility increased, the participants started paying less attention to the adaptive pursuit rotor task and more attention to the visual identification task. This was true when natural speech was presented too. However, the magnitude of change in average speed and reaction time for hits was higher for cell phone as compared to natural speech.

Hence, on the whole, when the participants were performing a consistent mapped task and when the speech presented was unpredictable, listening to cell phone speech or natural speech had a beneficial effect on visual word identification task. However, listening to cell phone speech or natural speech had a detrimental effect on the adaptive pursuit rotor task. The extent of detrimental/beneficial effects depended on the type of speech heard. Figure 41 shows the relationship between intelligibility and visual motor performance and reaction time for hits and visual motor performance for unpredictable sentences during consistent mapping condition

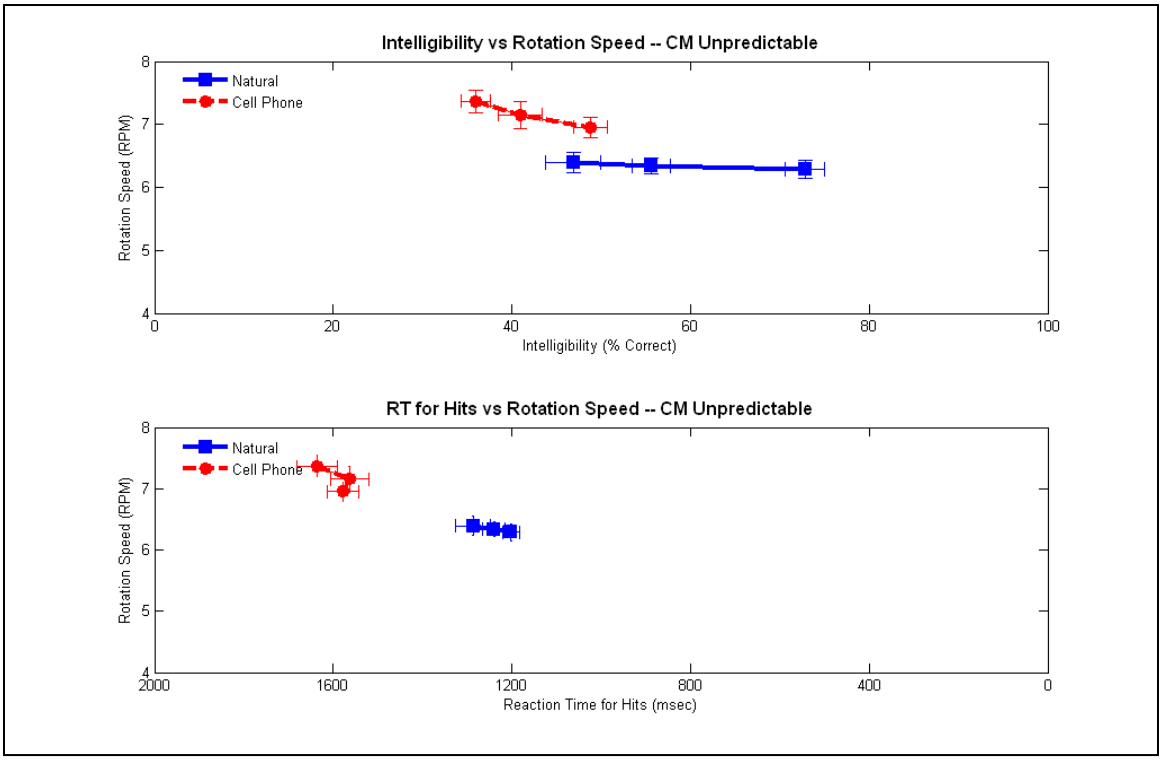

Figure 41. Relationship between intelligibility and visual motor performance and reaction time for hits and visual motor performance for unpredictable sentences during consistent mapping condition

When APR performance, speech intelligibility, and reaction time for hits in visual word identification task were compared for predictable sentences under varied mapping condition, it was found that when participants were listening to cell phone speech, increased intelligibility led to a decrease in average rotation speed and a decrease in reaction time for hits. This showed that at lower intelligibility levels, the participants paid more attention to the APR task and less attention to visual identification task. However, as intelligibility increased, the participants paid less

attention to the APR task and more attention to the visual identification task.

When natural speech was presented, increased intelligibility led to an increase in average rotation speed and an increase in reaction time for hits. This indicated the fact that as intelligibility of the sentences presented increased, the participants paid more attention to the adaptive pursuit rotor task and less attention to the visual identification task.

Hence, on the whole, when the participants were performing a varied mapped task and when the speech presented was predictable, listening to cell phone speech had a harmful effect on the performance on adaptive pursuit rotor task. Also, listening to cell phone speech had a beneficial effect on the visual identification task. Moreover, listening to natural speech had detrimental effects on visual word identification task and beneficial effects on the performance on the adaptive pursuit rotor task. Figure 42 shows the relationship between intelligibility and visual motor performance and reaction time for hits and visual motor performance for predictable sentences during varied mapping condition

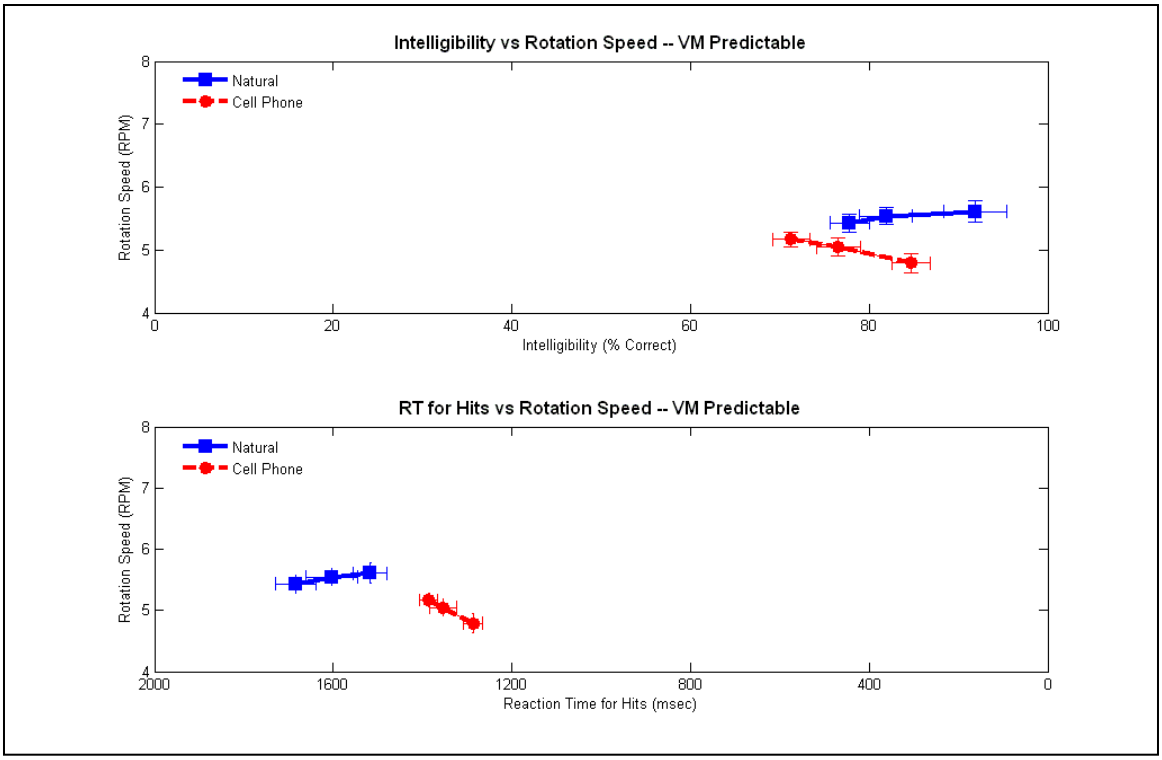

Figure 42. Relationship between intelligibility and visual motor performance and reaction time for hits and visual motor performance for predictable sentences during varied mapping condition

When APR performance, speech intelligibility, and reaction time for hits in visual word identification task were compared for unpredictable sentences in varied mapping condition, it was found that when participants were listening to cell phone speech, increased intelligibility led to a decrease in average rotation speed and a decrease in reaction time for hits. This showed that at lower intelligibility levels, the participants paid more attention to the APR task and less attention to the visual identification task. However, as the intelligibility increased, the participants started paying less attention to the adaptive pursuit rotor task and more attention to the visual

identification task. This was true when natural speech was presented too. However, the magnitude of change in average speed and reaction time for hits was higher for cell phone as compared to natural speech. This was true when natural speech was presented too. However, the magnitude of change in average speed and reaction time for hits was higher for cell phone as compared to natural speech. The reduction in average speed of rotation was higher when participants were listening to cell phone speech as compared to natural speech. However, the reduction in reaction time for hits was higher when participants were listening to natural speech as compared to cell phone speech.

Hence, on the whole, when the participants were performing a varied mapped task and when the speech presented was unpredictable, listening to cell phone speech or natural speech had a beneficial effect on visual word identification task. Also, listening to cell phone speech or natural speech had a detrimental effect on the adaptive pursuit rotor task. The extent of detrimental/beneficial effects depended on the type of speech heard. Figure 43 shows the relationship between intelligibility and visual motor performance and reaction time for hits and visual motor performance for unpredictable sentences during varied mapping condition

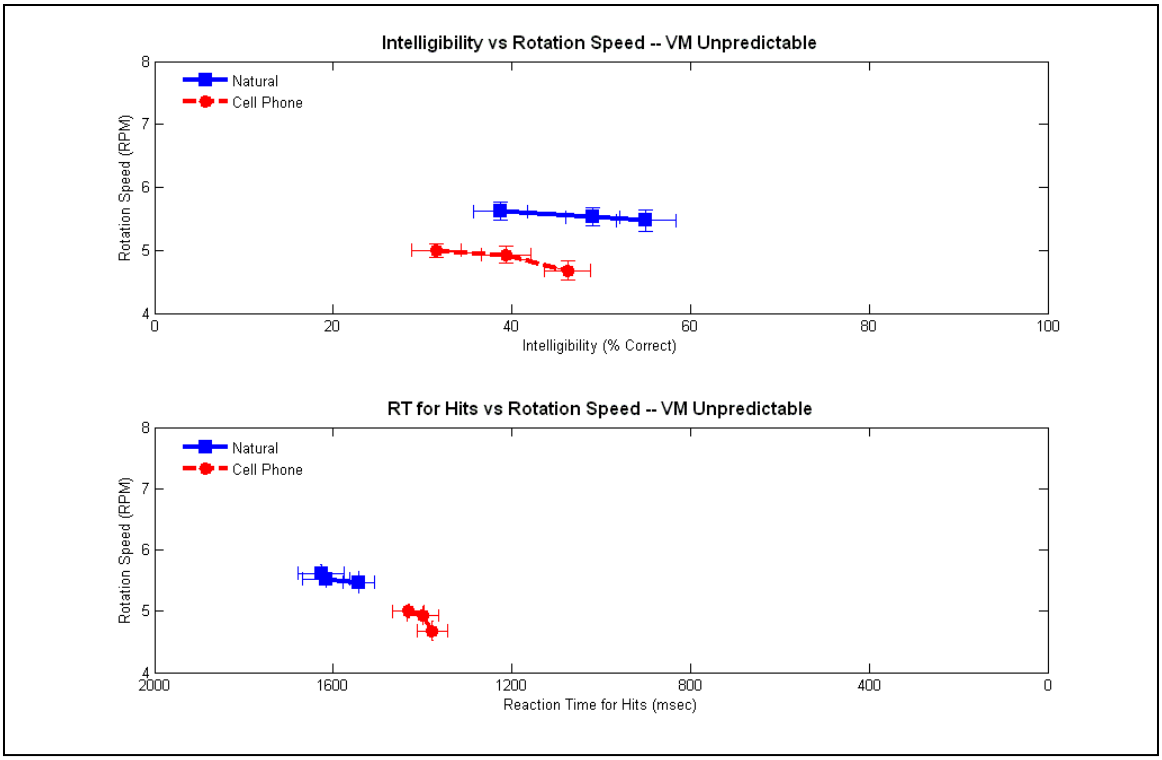

Figure 43. Relationship between intelligibility and visual motor performance and reaction time for hits and visual motor performance for unpredictable sentences during varied mapping condition

CHAPTER VII

Discussion

Experiment 1 & 2

The purpose of experiment 1 was to study the effects of cell phone speech, synthetic speech, and natural speech on speech intelligibility and performance on a simultaneous visual-motor task. It was expected that cell phone speech would have lower intelligibility as compared to synthetic speech and natural speech. Also, the speed of rotation during cell phone speech would be lower than during natural speech and synthetic speech. A dual-task methodology using an auditory word repetition task as the primary task and adaptive pursuit rotor task was developed and employed.

First, it was found that cell phone speech was less intelligible than synthetic speech. This pattern of results was consistent with the notion that the acoustic cues were poorly represented in the cell phone speech. Context helped cell phone speech more than synthetic speech. Also, the difference in intelligibility between predictable and unpredictable sentences was higher for cell phone speech as compared to synthetic speech. This pattern of results was consistent with the notion that the acoustic cues present in synthetic speech were properly represented. Furthermore, the time taken to repeat the last word of the sentence correctly was higher for cell phone speech as compared to synthetic speech. This was due to the fact that the intelligibility

of the cell phone speech was very low and the participants had to spend more time to process the cell phone speech.

Second, a dual-task procedure using adaptive pursuit rotor was effective at measuring increased processing demands in speech perception. It was found that as SNR increased the performance on simultaneous visual-motor task decreased for cell phone speech. However, this was not true for synthetic speech. This pattern of results might be because of the lower intelligibility associated with cell phone speech. At low intelligible levels of cell phone speech, the participants might have paid more attention to the visual-motor task as compared to the auditory word repetition task. This was verified by the low intelligibility scores, lower reaction time and higher mean speed of rotation for cell phone speech. Furthermore, for synthetic speech, the non-human nature of the speech forced the participants to pay more attention to the auditory task which was verified by higher intelligibility, higher reaction time and lower average speed. These results suggest that the use of a dual-task procedure may work well for measuring decreased signal quality for other signal simultaneous tasks as well. Therefore, the results of experiment 1 are methodologically important as well as directly relevant to measure the interference caused on speech perception due to various simultaneous tasks.

Though results of experiment 1 gave information about how participants performed on speech repetition task, it did not give a profile of

participants' performance across a range of SNRs. Also, the results did not give any information about the strategies used by the participants while listening to different qualities of speech and simultaneously performing a visual-motor task. Experiment 2 was designed to allow natural speech to be systematically compared with cell phone speech at different signal qualities.

First, it was found that as the signal quality improved the performance of visual-motor task decreased during cell phone speech. However, this decrement was found up to certain threshold in the SNR. After the threshold, the performance started to improve. This was due to the fact that at very low SNRs, the intelligibility of cell phone speech was very low and participants paid more attention to the visual-motor task. As the quality of the signal improved, the intelligibility improved and the participants paid less attention to the visual-motor task and hence there was a decrement in performance. After the threshold, the signal quality was so high that the participant's were able to pay more attention to the visual-motor task and hence there was an increase in performance. This was not true for natural speech. This was due to the fact that the intelligibility of the natural speech was high and the participants paid equal attention to both the tasks.

Second, it was found that, as expected, the intelligibility of natural speech was higher than that of cell phone speech even when natural speech was presented at much lower SNR. This pattern of results was consistent

with the notion that perceiving cell phone speech at noisier environments is much more difficult than perceiving natural speech at the same environment.

It was also found that cell phone speech and natural speech imposed different cognitive loads on the listener. Cell phone speech had a lower cognitive load than natural speech. This might be because of the fact that cell phone speech was presented at a much higher SNR compared to natural speech. Also, cell phones had a higher cognitive load compared to natural speech when the SNR levels were the same. This indicated the fact that the listeners had to spend more resource while listening to cell phone speech compared to natural speech. This was expected as the acoustic cues required for speech perception were represented clearly in natural speech as compared to cell phone speech. Moreover, as the signal got better, the cognitive load for listening to natural speech was far lower than listening to cell phone speech. This pattern of results was consistent with the notion that the listeners have to spend more resources while listening to cell phone speech as compared to natural speech.

The relationship between signal-to-noise ratio and intelligibility was linear in nature for cell phone speech ($R^2 = 0.934$). As expected, increased SNR led to an increase in intelligibility. The 50% crossover happened at around 4 dB SNR. However, the relationship between SNR and intelligibility for natural speech was logistical in nature ($R^2 = 0.966$). The 50% crossover happened at around -4 dB SNR. At low SNR levels for natural speech,

increased SNR did not improve intelligibility to a greater extent. In the regions of moderate SNR, increased SNR helped intelligibility quite a bit. However, at higher SNRs, increased SNR did not help intelligibility a lot.

There was a linear relationship between SNR and intelligibility for predictable and unpredictable cell phone speech. For predictable sentences, the 50% crossover happened at around 2 dB SNR ($R^2 = 0.947$) and for unpredictable sentences, the 50% crossover happened at around 8 dB SNR ($R^2 = 0.826$). Also, there was a logistical relationship between SNR and intelligibility for predictable and unpredictable natural speech. For predictable sentences, the 50% crossover happened at around -5 dB SNR ($R^2 = 0.972$) and for unpredictable sentences, the 50% crossover happened at around -3 dB SNR ($R^2 = 0.908$). As expected, semantic context of the sentences presented helped cell phone speech much better than natural speech to attain 50% intelligibility.

Experiment 3

A few studies used dual-task paradigms to measure listening effort (Broadbent, 1955; Downs, 1982; Strayer, Drews, & Johnston, 2003; Strayer & Burns, 2004). A decrease in performance on the secondary task has been related to the amount of effort expended in performing the primary task (Broadbent, 1955; Kahneman, 1973). The findings of the current study are similar to findings of Broadbent (1955), who used the simultaneous tracking task as the secondary task to compare filtered speech with frequency

transposed speech. A decrease in secondary tracking task was found for the frequency-transposed speech but not for filtered speech. However, no intelligibility difference was found between filtered speech and frequency-transposed speech. The implication of this study was that the importance of multiple criteria in the assessment of communication channels was suggested by their effect on their simultaneous tasks (Broadbent, 1955).

According to limited-channel capacity theories, there is one overall resource for attention capacity. Resources from the single capacity are divided according to the task-processing demands. Of processing load is greater than the overall capacity, performance on one of the tasks will decrease. So, when there is more than one task to process simultaneously, attention would be divided to process information or accomplish tasks. Adding a secondary task overloads the overall processing capacity, and with limited resources, the performance on the secondary task would decrease. Accordingly, there should be no effect on the performance in the primary task where as there should be a negative effect on the performance in the secondary task. However, the results of experiments 1 and 2 suggested different differential effects on speech perception task and adaptive pursuit rotor task at different levels of SNR.

According to multiple resource theory, humans do not have one information processing source that can be tapped, but different pools of resources that can be tapped simultaneously. Depending on the nature of the

task, these resources may have to process information sequentially if the different tasks require the same pool of resources, or can be processed in parallel if the task requires different resources. Wickens' theory viewed performance decrement as a shortage of these different resources and described human processing capabilities as having limited capacity for processing information simultaneously. Cognitive resources are limited and a supply and demand problem occurs when the individual performs two or more tasks that require a single resource. Excess workload caused by a task using the same resource as another task simultaneously can cause problems and would result in a lower performance. However, it should be noted that, when the modalities of the simultaneous tasks were different, there would not be a change in the performance of the task.

Even though the results of experiment 1 and 2 fit with Wickens' multiple resource theory to some extent, different detrimental effects for cell phone and natural speech led to the conclusion that there might be component of attention involved that need to be investigated deeper to better understand these results.

The purpose of experiment 3 was to study the effects of automatic and controlled processing tasks on cell phone and natural speech perception and simultaneous visual-motor tracking performance. Consistent Mapping (CM) and Varied Mapping (VM) tasks were used to experimentally induce automatic and controlled processing respectively. It was expected that the

speech intelligibility for cell phone and natural speech while doing a CM task would not be much different. However, when the task was a VM task, natural speech would have higher intelligibility as compared to cell phone speech. Also, it was expected that, the performance on the adaptive motor task would be better during CM task as compared to VM task during both cell phone and natural speech.

As expected, even though natural speech was presented at much lower SNR compared to cell phone speech, natural speech had a higher intelligibility. Also, intelligibility during consistent mapping was higher than varied mapping. This might be because of the fact that during consistent mapping, the target words were well learnt and more resources could be allocated to the auditory task. During visual mapping, since the target words were new, more resources had to be spent compared to CM task in order to complete the task.

Contrary to the initial assumption, it was found that the average speed of rotation during cell phone speech was higher when compared to natural speech when the visual word recognition task was consistently mapped. However, when the task was variably mapped, the average speed of rotation during natural speech was higher when compared to natural speech. This might be due to the fact that when the visual word identification task was consistently mapped, the task was well learnt and since the signal presented had less intelligibility, more resources could be allotted to the tracking task

thereby increasing the performance. When natural speech was presented, the signal was more intelligible and the resources had to be shared with both speech perception task and tracking task and eventually the performance on tracking task reduced. When the visual word identification task was variably mapped, even though the task was well learnt, the target words were different and hence more resources had to be spent in order to complete the task. The allocation of more resources for the word identification task reduced the overall available resources and hence performance on tracking task reduced. Lower intelligibility of cell phone speech combined with untrained target words in the visual identification task reduced the performance on the tracking task. However, since the intelligibility of natural speech was better compared to cell phone speech, the performance on the tracking task was better for natural speech compared to cell phone speech when the visual identification task was variably mapped.

It was found that when the visual identification task was consistently mapped, it was quicker to find the target words in visual identification task when the auditory signal was natural speech as compared to cell phone speech. However, when the visual identification task was variably mapped, it was quicker to find the target words in visual identification task when the auditory signal was cell phone speech as compared to natural speech. Higher intelligibility of natural speed added with consistent mapped task made it possible to quickly identify target words. However, when the task was

variably mapped, lower intelligibility of cell phone speech and lower average speed of rotation when cell phone speech was presented made it possible to spend more resources on the visual identification task and hence it was easier to identify the target words when the speech presented was cell phone speech. This pattern of results was true when there were no target words present in the visual word recognition task.

One interesting finding from experiment 3 was about the way in which the participants approached the presence or absent of the target words. When the visual identification task was consistently mapped, the listeners behaved ideally while responding to whether the target words were present or absent. However, when the task was variably mapped, the listeners were more liberal – more likely to respond target word present when the target word was actually absent. Since automaticity for the new target words would not have developed while target words were variably mapped, the listeners had to look at all the words in the visual stimuli and respond whether target word was present or not there by making them more liberal.

These results might give an illusion that it is safer to talk on a cell phone and drive a car simultaneously when the driving map was consistently mapped. Driving on a familiar route is an example of consistently mapped task. However, it should be kept in mind that the traffic scenarios vary from time to time and hence we cannot call the task to be consistently mapped. Since the responses to different traffic scenarios at different conditions are

different, the task is predominantly a varied mapping task with a small percentage of automatic component built into it.

CHAPTER VIII

Conclusions

Dual-task Paradigms and Listening Effort

The use of intelligibility measures alone for the evaluation of different types of speech in complex environments has provided inconsistent findings in a variety of experimental paradigms. A supplemental method of evaluation was proposed in this study. It was expected that the additional measures might provide additional information about the effort required to perceive different kinds of speech in noise in complex listening environments. Furthermore, it was expected that this additional information would assist our understanding of how attentional mechanisms affect speech perception of different kinds of speech in complex listening environments in a manner that might not be completely reflected by intelligibility measures alone. In this study, two dual-task paradigms were developed to measure the changes in listening effort based on attention and capacity theories (Broadbent, 1958; Kahneman, 1973, Pisoni, 1982; Schneider & Shiffrin, 1977; Wickens, 1984). It was expected that an increase in processing demands in a speech signal by the addition of noise would result in decrease in the secondary task due to limited capacity as indication of increased listening effort whereas speech intelligibility measures do not reflect the change in signal quality because of the listening effort. It was also hypothesized that increase in listening effort

during cell phone speech would have a higher detrimental effect on the simultaneous task performance compared to the increase in listening effort during natural speech because of the manner in which the different acoustic cues are present in these two signals. Finally, it was hypothesized that varied mapping would produce a greater reduction in performance as compared to consistent mapping while listening to cell phone speech as opposed to natural speech. Therefore, differences in intelligibility of cell phone speech and natural speech would be better quantified in more complex listening environments rather than in controlled laboratory environments. Three experiments were conducted to test these hypotheses.

The overall results of this study demonstrated the usefulness of dual-tasks in measuring listening effort and intelligibility during the perception of cell phone speech and natural speech in complex listening environments. The results suggested that intelligibility of speech affected the performance on simultaneous visual-motor task much more for cell phone speech that natural speech. Also, the quality of the cell phone speech presented was much more important for performance of simultaneous task as compared to the quality of the natural speech presented. The results also suggested that the type of mapping of the secondary task had differential effects for cell phone speech and natural speech. While listening to cell phone speech and performing simultaneous tasks, consistent mapping was required for at least one of the tasks to perform both the tasks better. However, type of mapping did not play

any role in listening to natural speech and performing simultaneous tasks together. The results also suggested an abandonment of auditory word repetition task to do better in the simultaneous visual-motor task at lower intelligibilities while listening to cell phone speech. However, this was not true while the participants were listening to natural speech. These results indicated a differential change in listening effort while listening to cell phone speech and natural speech. The change in effort was reflected in intelligibility and simultaneous visual-motor task performances. The current study partially supported the proposed hypotheses and the current theories of attention by demonstrating both the limitations of our mental capacity in simultaneous processing of multiple tasks and successful processing of linguistic information when two different modalities were involved. This pattern of results suggested that careful implementation of dual-task paradigms is required for an efficient measure of speech intelligibility in complex listening environments and of listening effort. More work needs to be conducted to determine the usefulness of dual-task paradigms as an underlying standard protocol for the evaluation of speech intelligibility of different types of speech in complex listening environments.

The task load measured using the subjective method like the NASA-TLX used here were useful for getting the information about the cognitive load imposed on the participant while doing the task after the task has been completed. The dual-task method proposed here collected the cognitive load

imposed while the participant was doing the experiment. This method was found to be more reliable compared to the subjective NASA-TLX in measuring the load while listening and performing simultaneous task. This was evident from the significance values (p-values) for intelligibility and average speed and intelligibility and NASA-TLX correlations.

In the future, the following questions need to be answered. First, the difficulty of the visual-motor task should be equated approximately to the same difficulty level as the visual identification task. These results were obtained from listeners who had more practice time with the visual word identification task as compared to adaptive pursuit rotor task. Though more time with visual word identification task is attributed to the development of consistent mapping, differences in practice could also have caused these measurements. Hence, in future, these tests should be administered on people who had same amount of practice on visual word identification task and adaptive pursuit rotor task. Next, intelligibility with the visual word identification task alone should be compared with adaptive pursuit rotor task to determine whether the participants really gave up on the adaptive pursuit rotor task at lower intelligibility levels in Experiment 2. Third, a consistent mapping should be developed for auditory stimulus to determine whether the differences in input modality have any effect on the outcome. Fourth, cell phone speech should be presented at much higher levels of signal-to-noise ratios (greater than 10 dB) to further characterize the non-linear relationship

between intelligibility and adaptive pursuit rotor performance. Finally, additional levels of distractions should be investigated to find performance functions for simultaneous tasks. Above all, the effectiveness of dual-task paradigms using adaptive pursuit rotor task and visual word identification task needs to be determined by replicating the results with other types of speech (e.g., synthetic speech, hearing aid speech, reduced channel speech, distorted speech) for measuring listening effort.

Intelligibility and Types of Speech

Natural speech had higher intelligibility compared to cell phone speech even though cell phone speech was presented at much higher SNRs as opposed to natural speech. This clearly shows that the perception mechanism for cell phone speech and natural speech are totally different. At lower intelligibility for cell phone speech, the participants abandoned the listening task and paid more attention to the visual-motor task. This was evident from higher speed or rotation at lower intelligibility levels for cell phone speech. However, for natural speech, there was no change visual-motor performance at lower intelligibility levels. It should also be noted that cell phone speech and natural speech were presented at different SNRs. This clearly shows the presence of different perceptive mechanisms for cell phone and natural speech. Moreover, context of the presented stimuli had differential effects on speech intelligibility of cell phone speech and natural speech. Differences between the intelligibility of predictable and unpredictable sentences was

higher for cell phone speech as compared to natural speech. Also, predictable sentences had higher intelligibility and lower visual-motor performance for cell phone speech.

At higher intelligibility levels (intelligibility greater than 75%), for cell phone speech, increase in intelligibility reduces the visual-motor performance. However for natural speech, there was no change in visual-motor performance for any change in intelligibility. At acceptable intelligibility levels (intelligibility greater than 85%), simultaneous visual-motor performance was higher for natural speech as compared to cell phone speech. As the cell phone intelligibility increases beyond 90 %, there was a rapid increase in average rotation speed. This rapid increase in average rotation speed happens when intelligibility crosses 95% for natural speech. Based on the pattern of results obtained, it could be concluded that cell phone speech intelligibility has asymptoted at 95% where as natural speech has the ability to be 99.9% intelligible. Overall, at higher intelligibility levels, cell phone speech would not have the same visual-motor performance as natural speech even though it had at lower intelligibility levels. At lower intelligibility levels, the higher performance on the visual-motor task was at the expense of auditory word repetition task.

The intelligibility data presented here were collected from undergraduate listeners with normal hearing. It is unclear whether the same set of results would be obtained with different groups of listeners (e.g., elderly

listeners, listeners with hearing impairment). Hence, more research should be conducted to generalize the effort required to perceive different kinds of speech in noise in complex listening environments. Most of the listeners volunteered in this study had prior experience of talking on a cell phone and driving simultaneously (29 out of 36) which is similar to listening to cell phone speech while doing adaptive tracking task. Hence, these results could not be used to generalize speech perception in complex listening environments for listeners without prior experience in dual-tasking.

Additional measures must be developed to effectively measure listening effort. When attention was measured using consistent mapping and varied mapping techniques, participants developed different strategies to do three tasks (auditory word repetition, visual-motor tracking, and visual word identification) simultaneously. During CM condition, the participants did not have any difficulty in performing three tasks simultaneously as the automaticity was developed for one of the tasks (visual word identification). However, when visual word identification was variably mapped, the participants devised different strategies to allocate resources for the three tasks so that all the three tasks could be performed simultaneously. Hence a more controlled experiment where in all the participants use the same resource allocation strategies should be conducted before generalizing these results. Also, the voice quality of the verbal response during the auditory word repetition task changed (e.g., shorter word durations while listening to

natural speech as compared to longer word duration while listening to cell phone speech, slurred articulation, addition or deletion or substitution of fricatives while listening to cell phone speech). This change in the quality of the vocal responses may have been resulted from different processing demands for cell phone speech and natural speech. This indicates that different types of speech not only affect perception, but eventual speech production. Furthermore, it suggests the importance of examining multiple dependent measures to evaluate the quality of speech.

In conclusion, based upon the results from the first experiment, the use of dual-task paradigms for the evaluation of different types of speech in complex environments is suggested. The second experiment showed that adverse listening environments have differential effects on cell phone speech and natural speech perception. Also, it proved the existence of a non-monotonic function relating intelligibility and simultaneous task performance. The third experiment showed that for an accurate measure of the evaluation of different types of speech in complex environments using dual-tasks; a careful calibration of the simultaneous tasks is required to make sure the overall task demands overload the overall attention capacity, which is limited in nature.

References

Alm, H., & Nilsson, L. (1994). Changes in driver behavior as a function of handsfree mobile phones – A simulator study. *Accident Analysis & Prevention*, 26 (4), 441 – 451.

Alm, H., & Nilsson, L. (1995). The effects of a mobile telephone task on driver behavior in a car following situation. *Accident Analysis & Prevention*, 27 (5), 707 – 715.

Assman, P., & Summerfield, Q. (2004). The perception of speech under adverse conditions. In Greenberg, S., Ainsworth, W. A., Popper, A. N., & Fay, R. R. (Eds.). *Speech processing in the auditory system* (p. 231 – 308). New York: Springer Verlag.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A proposed system and its control processes. In Spence, K. W. & Spence, J. T. (Eds.). *The psychology of learning and motivation: Advances in research and theory* (Vol. 2). New York: Academic Press.

Bargh, J. A. (1992). The ecology of automaticity: Toward establishing the conditions needed to produce automatic processing effects. *American Journal of Psychology*. 105 (2), 181 – 199.

Bilger, R.C., Nuetzel, J.M., Rabinowitz, W.M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research*. 27, 32 – 48.

Boothroyd, A. (1982). Communication aids for the deaf. In Bess, F. H., Freeman, B. A., & Sinclair, J. S. (Eds.), *Technology for independent living*, Washington: American Association for the Advancement of Science.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, 84 (1), 101 – 114.

Bower, G. H. (1967). A multicomponent theory of the memory trace. . In Spence, K. W. & Spence, J. T. (Eds.). *The psychology of learning and motivation* (Vol. 1, p. 230 - 235). New York: Academic Press.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Breslau, L, Estrin, D., Floyd, S., Heidemann, J., Helmy, A., Huang, P., McCanne, S., Varadhan, K., xu, Y., & Yu, H. (2000). Advances in network simulation. *IEEE Computer*, 33 (5), 59 – 67.

Briem, V., & Hedman, L. R. (1995). Behavioral effects of mobile telephone use during simulated driving. *Ergonomics*, 38 (12), 2536 – 2562.

Broadbent, D. E. (1952). Listening to one of two synchronous messages. *Journal of Experimental Psychology*, 44, 51 – 55.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon Press.

Brookhuis, K. A., De Vries, G., & De Waard, D. (1991). The effects of mobile telephoning on driving performance. *Accident analysis and Prevention*, 23, 309 – 316.

Brown, I. D., & Poulton, E. C. (1961). Measuring the spare 'mental capacity' of car drivers by a subsidiary task. *Ergonomics*, 4, 35 – 40.

Brown, I. D., Tickner, A. H., & Simmonds, D. C. V. (1969). Interference between concurrent tasks of driving and telephoning. *Journal of Applied Psychology*. 43, 462 – 482.

Brungart, D. S. (2001). Evaluation of speech intelligibility with the coordinate response measure. *Journal of the Acoustical Society of America*, 109 (5), 2276 – 2279.

Carrell, T. D., & Opie, J. M. (1992). The effect of amplitude comodulation on auditory object formation in sentence perception. *Perception & psychophysics*, 52(4), 437 – 445.

Charlton. S. G. (1996). Mental workload test and evaluation. In O'Brien, T. G., & Charlton, S. G. (Eds.). *Handbook of human factors testing and evaluation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chen, T., & Rao, R. R. (1998). Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86 (5), 837 – 852.

Choi, S. (2004). *The effect of compression on speech perception as reflected by attention and intelligibility measures*. Unpublished doctoral dissertation, University of Nebraska – Lincoln, Lincoln, NE.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In Cole, R. A. (Ed.), *Perception and production of fluent speech*, (pp. 133 – 163). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Davis, C., & Davis, D. (1997). *Sound system engineering*. Newton, MA : Focal Press.

Diehl, R. L., Lotto, A. J., & Holt, H. L. (2004). Speech perception. *Annual Review of Psychology*, 55, 149 – 179.

Ding, X., Erickson, T., Kellogg, W. A., Levy, S., Christensen, J. E., Sussman, J., Wolf, V. T., & Bennett, W. E. (2007). An empirical study of the use of visually enhanced VoIP audio conferencing: The case of IEAC. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, San Jose, CA. 1019 – 1028

Dirks, D. D., & Bower, D. R. (1969). Masking effects of competing messages. *Journal of Speech and Hearing Research*, 12, 229 – 245.

Dirks, D. D., Morgan, D. E., & Dubno, J. R. (1982). A procedure for quantifying the effects of noise on speech recognition. *Journal of Speech and Hearing Research*, 47, 114 – 123.

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, 30, 412 – 430.

Douskalis, B. (2000). IP Telephony: *The Integration of Robust VoIP Services*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Downs, D. W., & Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech and Hearing Research*, 21, 702 – 714.

Drews, F. A., & Strayer, D. L. (2004). Profiles in driver distraction: Effects of cell phone conversation on younger and older drivers. *Human Factors*, 46 (4), 640 – 649.

Edwards, B. W. (2000). Beyond amplification: Speech processing techniques for improving speech intelligibility in noise with hearing aids. *Seminars in Hearing*, 21 (2), 137 – 156.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review*, 89, 627 – 661.

Elliott, L. L. (1995). Verbal auditory closure and the speech perception in noise (SPIN) test. *Journal of Speech and Hearing Research*, 38, 1363 – 1376.

Ellis, D. P .W (1996). *Prediction-driven computational auditory scene analysis*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge , MA.

Elman, J., & McClelland, J. (1986). Exploiting lawful variability in speech wave. In Perkell, J. S., & Klatt, D. H. (Eds.). *Invariance and variability in speech process* (p. 360 – 380). Hillsdale, NJ: Lawrence Erlbaum Associates.

Endsley, M. (1995). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting,* 97 – 101.

Fairclough, S. H., Ashby, M. C., Ross, T., & Parkes, A. (1991). Effects of handsfree telephone use on driving behavior. In *Proceedings of the 24th ISATA International Symposium on automotive technology and automation*, Croydon, England. 403 – 409.

Fisk, A. D., & Schneider, W. (1983). Category and Word Search: Generalizing search principles to complex processing. *Journal of Experimental Psychology. Learning Memory, and Cognition,* 9 (2), 177 – 195.

Fisk, A. D., & Schneider, W. (1984). Memory as a function of attention, level of processing, and automatization. *Journal of Experimental Psychology. Learning Memory, and Cognition,* 10 (2), 181 – 197.

Francis, A. L., Nusbaum, H. C., & Fenn, K. (2007). Effects of training on the acoustic-phonetic representation of synthetic speech. *Journal of Speech, Language, and Hearing Research,* 50, 1445 –1465.

Fucci, D., Reynolds, M. E., Bettagere, R., & Gonzales, M. D. (1995). Synthetic speech intelligibility under several experimental conditions. *Augmentative and Alternate communication*, 11, 113 – 117.

Gatehouse. S., & Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *British journal of Audiology,* 24, 63 – 68.

Giolas, T. G., & Epstein, A. (1963). Comparative intelligibility of word lists and continuous discourse. *The Journal of Speech and Hearing Research*, 12, 349 – 358.

Gong, L., & Lai, J. (2001). Shall we mix synthetic speech and human speech? Impact on users' performance, perception and attitude. *Proceedings of the SIGCHI conference on human factors in computing systems*. Seattle, Washington, USA. 158 – 165.

Goodman, M. F., Bents, F. D., Tijerina, L., Wierwille, W. W., Lemer, N. A. & Benel, D. (1997). *An investigation on the safety implications of wireless communications in vehicles. Report Summary*. National Highway Traffic Safety Administration (NHTSA), US department of Transportation, Washington, DC.

Goodman, M.F., Tijerina, L., Bents, F.D., & Wierwille, W.W. (1999). Using cell phones in vehicle: Safe or unsafe? *Transportation Human Factors, 1*, 3-42.

Greenspan, S. L., Bennett, R. W., & Syrdal. A. K. (1998). An evaluation of the diagnostic rhyme test. *International Journal of Speech Technology*, 2 (3), 201 – 214.

Gustafsson, H. A., & Arlinger, S. D. (1991). Masking of speech by amplitude-modulated noise. *Journal of the Acoustical Society of America*, 151 (3), 441 – 445.

Hart, S. G., & Staveland, L. E. (1988). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Hancock, P. S., & Meshkati, N. (Eds.). *Human Mental Workload* (p. 139 – 183). Amsterdam: North-Holland.

Hawkins, D. B., & Yacullo, W. S. (1984). Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation. *Journal of Speech and Hearing Disorders*, 49, 278 – 286.

Hawley. M. E. (Ed.). (1977). *Speech intelligibility and speaker recognition* Stroudsburg, PA: Dowden, Hutchinson, & Ross, Inc.

Helfer, K. S. (1991). Everyday speech understanding by older listeners. *Journal of the Academy of Rehabilitative Audiology*, 24, 17 – 34.

Helfer, K. S., & Wilber, L. A. (1990). Hearing loss. Aging, and speech perception in reverberation and noise. *Journal of Speech and Hearing Research*, 33, 149 – 155.

Hoffman, L., Yang, X., Bovaird, J. A., & Embretson, S. E. (2006). Measuring attentional ability in older adults. *Educational and Psychological Measurement*, 66 (6), 984 – 1000.

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J., & Brown, J. (2006). Driver distraction. The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis & Prevention*, 38, 185 – 191.

Irwin, M., Fitzgerald, C., & Berg, W. P. (2000). Effects of the intensity of wireless telephone conversations on reaction time in braking response. *Perceptual and Motor Skills*. 90, 1130 – 1134.

James. W. (1980). *The principles of Psychology*. New York: Holt.

Jamieson, D. G., Deng, L., Price, M., Parsa, V., & Till, J. (1996). Interaction of speech disorders with speech coders: Effects on speech intelligibility. In *Fourth International Conference on Spoken Language*, 2, 737 – 740.

Jerger, D., & Jerger. J. (1983). Evaluation of diagnostic audiometric tests. *Audiology*, 22, 144 – 161.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Jusczyk, P. W., & Luce, P. A. (2002). Speech perception and spoken word recognition: past and present. *Ear & Hearing*, 23 (1), 2 – 40.

Kahneman, D. (1973). *Attention and Effort*. Englewood cliffs, NJ. Prentice Hall.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61 (5), 1337 – 1351.

Kent, R. D., & Read, C. (2002). *The acoustic analysis of speech*. San Diego, CA: Singular Publishing Group, Inc.

Klatt, D. H. (1978). SCRIBER and LAFS: Two new approaches to speech analysis. In Lea, W.A. (Ed.), *Trends in speech recognition*. Englewood Cliffs, NJ: Prentice hall

Kubose. T. T., Bock, K., Dell, G.S., Garnsey, S. M., Kramer, A. F., & Mayhugh, J. (2006). The effects of speech production and speech comprehension on simulated driving performance. *Applied Cognitive Psychology*. 1, 43 – 63.

Kurose, J. F., & Ross, K. W. (2004). *Computer Networking: A Top-Down Approach Featuring the Internet*. Adisson Wesey.

Lai, J., Wood, D., & Considine, M. (2000). The effect of task conditions in the comprehensibility of synthetic speech. *The proceedings of CHI '00*, ACM Press, 321 – 328.

Lam, K.H., Ay, O. C., Chan, C. C., Hui, K. F., & Lau, S. F. (1996). Objective speech quality measure for cellular phone. *Proceedings of IEEE internation conference on Acoustics, Speech and Singal Processing*, Vol. 1, Atlanta, GA, May 1996. 487 – 490.

Lamble, D., Kauranen, T., Laakso, M., & Summala, H. (1999). Cognitive load and detection thresholds in car following situations: Safety implications for using mobile telephones while driving. *Accident Analysis & Prevention*, 31 (6), 617 – 623.

Larsby, B., & Arlinger, S. (1994). Speech recognition and just-follow conversation tasks for normal hearing and hearing impaired listeners with different maskers. *Audiology*, 33, 165 – 176.

Lee, J. D., Vaven, B., Haake, S., & Brown, T. L. (2001). Speech-based interaction with in-vehicle computers: The effects of speech-based e-mail on drivers' attention to the roadway. *Human Factors*, 43, 631 – 640.

Lavandier, M., & Cilling, J. F. (2010). Prediction of binaural speech intelligibility against noise in rooms. *The Journal of the Acoustical Society of America*, 127 (1), 387 – 399.

Liberman, A. M., Cooper, F. S., Shanwalker, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431 – 461.

Lindbloom, B. (1990). On the communication process: Speaker-listener interaction and development of speech. *Augmentative and Alternate Communication*, 6 (4), 220 – 230.

Logan, G. D. (1992). Attention and preattention theories of automacity. *American journal of psychology*, 105 (2), 523 – 553.

Logan, J. S., Greene, B. G., & Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *The Journal of the Acoustical Society of America*, 86 (2), 566 – 581.

Longworth-Reed, L., Brandewie, E., & Zahorik, P. (2009). Time-forward speech intelligibility in time-reversed rooms. *The Journal of the Acoustical Society of America*, 125 (1), 13 – 19.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 1983, 25 (10, 17 – 32.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics*, 62 (3), 612 – 625.

Luximon, A., & Goonetilleke, R. S. (2001). Simplified subjective workload assessment technique. *Ergonomics*, 44 (3), 229 – 243.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1 – 36.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. in Altmann, G. (Ed.), *Cognitive Models of Speech Perception: Psycholinguistic and Computational Perspective* (122 – 147). Cambridge, MA: MIT Press

Marsden-Johnson, J. (1990). *Perception of synthetic speech during dual task performance*. Doctoral Dissertation, Northwestern University.

Marslen-Wilson, W. (1987). Functional Parallelism in spoken word-recognition. *Cognition*. 25, 71 – 102.

Marslen-Wilson, W. (Ed.). (1989). *Lexical representation and process.* Cambridge, MA: MIT Press

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8, 1 – 71.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10, 29 – 63.

Markopoulou, A. M., Tobagi, F. A., & Karam, M. J. (2002). Assessment of VoIP quality over internet backbones. Infocom 2002. *Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. 1, 150 – 159.

Martin, J. (1976). *Telecommunications and the comput.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Matthews, R., Legg, S., & Charlton, S. (2003). The effects of cell phone types on drivers' subjective workload during concurrent driving and conversing. *Accident Analysis and Prevention*. 35 – 451 – 457.

McCarthy, G., & Donchin, E. (1981). A metric for thought: A comparison of P300 latency and reaction time. *Science*, 211, 77 – 79.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1 – 86.

McGurk, H., & MacDonald. J. (1976). Hearing lips and seeing voices. *Nature,* 264. 746 – 748.

McKnight, A. J., & McKnight, S. J. (1993). The effect of cellular phone use upon driver attention. *Accident Analysis and Prevention*, 25 (3), 259 – 165.

Miller, G. A. (1947). Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *Journal of the Acoustical Society of America*, 191, 609 – 619.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41, 329 – 335.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27, 338 – 352.

Mirenda, P., & Beukelman, D. R (1990). A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. *Augmentative and Alternate Communication*, 6 (1), 61 – 68

Moore, B. (1998). *Cochlear Hearing Loss*. Whurr, London: John Wiley & Sons, Inc.

Moray, N. (1967). Where is attention limited? A survey and a model. *Acta Psychologica*, 27, 84 – 92.

Nabelek, A. K., & Letowski, T. R. (1985). Vowel confusions of hearing impaired listeners under reverberant and nonreverberant conditions. *Journal of Speech and Hearing Research*, 50, 126 – 131.

Nabelek, A. K., & Mason, D. (1981). Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms. *Journal of Speech and Hearing Research*, 24, 375 – 383.

Nabelek, A. K., & Picket, J. M. (1974). Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing impaired listeners. *Journal of Speech and Hearing Research*, 17, 724 – 739.

Nabelek, A. K., & Robinette, L. (1978). Reverberation as a parameter in clinical testing. *Audiology*, 17, 239 – 259.

Narbutt, M., Kelly, A., Murphy, L., & Perry, P. (2005). Adaptive VoIP playout scheduling: Assessing user satisfaction. *IEEE Internet computing*, 9 (4), 28 – 34.

Navon, D., Gopher, D. (1979). On the economy of the human information processing system. *Psychological Review*, 86, 214 – 255.

Nilsson, M., Sigfrid, S. D., & Sullivan, J. A. (1994). Development of hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95 (2), 1085 – 1099.

Noguchi, T., Demura, S., Nagasawa, Y., & Uchiyama, M. (2005). The practice effect and its difference of the pursuit rotor test with dominant and non-dominant hands. *Journal of Physiological Anthropology and Applied Human Science*. 24 (6), 589 – 593.

Nooteboom, S G. (1983). The temporal organization of speech and the process of spoken-word recognition. *IPO Annual Progress Report*, 18, 32 – 36.

Norman, D. A., & Babrow, D. G. (1975). On data-limited and resource limited processes. *Cognitive Psychology*, 7, 44 – 65.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, 33 (4), 17 – 33.

Parasuraman, R. (1990). Event-related brain potentials and human factors research. In Rohrbaugh, J. W., Parasuraman, R., & Johnson Jr, R. (Eds.), *Event-related Brain Potentials: Basic Issues and Applications* (279 – 300). New York : Oxford University Press.

Parkes, A., & Hooijmeijer, V. (2000). *The influence of the use of mobile phones on driver situation awareness*. Retrieved from http://www-nrd.nhtsa.dot.gov/departments/Human%20Factors/driver-distraction/PDF/2.PDF

Parkes, A. & Hooijmeijer, V. (2000). *The influence of the use of mobile phones on driver situation awareness*. Available at http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/PDF/2.PDF

Patten, C. J. D., Kircher, A., Ostlund, J., & Nilsson, L. (2004). Using mobile phones: cognitive workload and attention resource allocation. *Accident Analysis & Prevention*, 36 (3), 341 – 350.

Pichney, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing: 1. Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, 28, 96 – 103.

Pike, R., McFarland, K., & Dalgleish, L. (1974). Speed-accuracy tradeoff models for auditory detection with deadlines. *Acta Psychologica*, 38, 379 – 399.

Pisoni, D. B. (1982). Perception of speech: The human listener as a cognitive interface. *Speech Technology*, 10 – 23.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, 13 (1-2), 109 – 125.

Pisoni, D. B., Nusbaum, H. C., & Greene, B. G. (1985). Perception of synthetic speech generated by rule. In *Proceedings of IEEE*, 73 (11), 1665 – 1676.

Pisoni, D. B., & Tash, J. (1982). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, 15 (2), 285 – 290.

Posner, M. I. (1978). *Chronometric explorations of mind: the third Paul. M. Fitts lectures delivered at the University of Michigan, September, 1976*. New York: The Halsted Press

Pratt, R. L. (1981). On the use of reaction time as a measure of intelligibility. *British Journal of Audiology*, 15, 253 – 255.

Raghavendra, P., & Allen, G. D. (1993). Comprehension of synthetic speech with three text-to-speech systems using a sentence verification paradigm. *Augmentative and Alternate Communication*, 9 (2), 126 – 133.

Rakauskas, M. E., Gugerty, L. J., & Ward. N. J. (2004). Effects of naturalistic cell phone conversations on driving performance. *Journal of Safety Research*. 35, 453 – 464.

Rappaport, T. S. (1996). *Wireless Communications: Principles and Practice.* New Jersey: Prentice-Hall, Inc.

Recarte, M. A., & Numes, L. M. (2003). Mental workload while driving: Effects on visual search, discrimination and decision making. *Journal of Experimental Psychology: Applied*, 9 (2), 119 – 137.

Redelmeier, D. A., & Tibshirani, R. J. (1997). Association between cell phone calls and motor vehicle collisions. *The New England Journal of Medicine*, 336 (7), 453 – 458.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, 81, 275 – 280.

Reid, G. B., Shingledecker, C. A., & Effemeier, F. T. (1981). Application of conjoint measurement to workload scale development. *Proceedings of the Human Factors Society*, 25, 522 – 526.

Reynolds, M. E., Bond, Z. S., & Fucci, D. (1996). Synthetic speech intelligibility: Comparison of native and non-native speakers of English. *Augmentative and Alternate Communication*, 12, 32 – 36.

Rolfe, J. M., & Lindsay, S. J. E. (1973). Flight deck environment and pilot workload: Biological measures of workload. *Applied Ergonomics*, 4, 199 – 206.

Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and Workload Profile. *Applied Psychology: An International Review*, 53 (1), 61 – 86.

Ryu, K., & Myung, R. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, 35 (2005), 991 – 1009.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, 110, 474 – 494.

Santen, J. P. H. V., Perceptual experiments for diagnostic testing of text-to-speech systems (1993). *Computer Speech and Language*, 7, 49 – 100.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America*. 35 (2), 200 – 206.

Sawusch, J. R. (1996) Instrumentation and Methodology for the study of speech perception. In N. J. Lass (Ed.), *Principles of Experimental Phonetics* (pp. 525 – 550). St. Louis, MO : Mosby

Scherz, J. W., & Beer, M. M. (1995). Factors affecting the intelligibility of synthesized speech. *Augmentative and Alternate Communication*, 11, 74 – 78.

Schneider, W., Dumais, S. T., & Shiffrin, R. M. (1984). Automatic and controlled processing in attention. In Parasuraman, R., & Davies, D. R. (Eds.). *Varieties of attention*. Orlando, FL: Academic Press.

Schneider, W., & Fisk, A. D. (1982). Degree of consistent training: Improvements in search performance and automatic process development. *Perception and Psychophysics,* 32 (2), 160 – 168.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing. I. Detection, search, and attention. *Psychological Review,* 84(1), 1 – 66.

Schroeder, M. R., & Atal. B. S. (1985). Code-Excited Linear Prediction (CELP): High quality speech at very low bit rates. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 10, 937 – 940.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors,* 24 (4), 395 – 408.

Shacham, N., Craighill, E. J., & Poggio, A. A. (1983). Speech transport in packet-radio networks with mobile nodes. *IEEE Journal on Selected Areas in Communications*, 1 (6), 1084 – 1097.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing. II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127 – 190.

Silverman, K., Kalyanswamy, A., Silverman, J., Basson, S., & Yashchin, D. (1993). Synthesizer intelligibility in the context of a name-and-address information service. *Third European Conference on Speech Communication and Technology*, 2169 – 2172.

Snoddy, G. S. (1926). Learning and stability: A psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology*, 10(1), 1 – 36.

Sonnenschein, S. (1985). The development of referential communication skills: Some situations in which speakers five redundant messages. *Journal of Psycholinguistic Research*, 12 (5), 489 – 508.

Sperry, J. L., Wiley, T.L., & Chial, M.R. (1997). Word recognition performance in various background competitors. *Journal of the American Academy of Audiology*, 8, 71 – 80.

Spiegel, M. F., Altom, A. J., Macchi, M. J., & Wallace, K. L. (1990). Comprehensive assessment of the telephone intelligibility of

synthesized and natural speech. *Speech Communication*, 9 (4), 279 – 291.

Spragins, J. D., Hammond, J. L., & Pawlikowski, K. (1991). *Telecommunications: Protocols and Design*. Reading, MA: Adisson-Wesley.

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech and Language*, 19, 96 – 104.

Strayer, D. L., & Drews. F. A. (2004). Profiles in driver distraction: Effects of cell phone conversations on younger and older drivers. *Human Factors*, 46, 640 – 649.

Strayer, D. L., Drews, F. A., Crouch, D. J., & Johnston, W. A (2005). Why do cell phone conversations interfere with driving? In Walker, W. R., & Herrmann, D. (Eds.), *Cognitive Technology: Transforming Thought and Society*. Jefferson, NC: McFarland & Company Inc.,

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, 9 (1), 23 – 32.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual task studies of simulated driving and conversing on a cell phone. *Psychological Science*, 12 (6), 462 – 466.

Syrdal, A. K., & Sciacca, B. A. (1994). *Testing the intelligibility of text-to-speech output with the diagnostic pairs sentence intelligibility evaluation*. Unpublished manuscript.

Tan, X., Wänstedt, S., & Heikkilä, G. (2001). Experiments and modeling of perceived speech quality samples. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2, 849 – 852.

Tanenbaum, A. (1996). *Computer Networks*, fourth edition. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Tarawaneh, M. S. (1991). *A conceptual model of driving task to evaluate measures of improving safety of elderly drivers.* Unpublished doctoral dissertation, University of Nebraska – Lincoln, Lincoln, NE.

Teder, H. (1990). Noise and speech levels in noisy environments. *Hearing Instruments*, 41, 32 – 33.

Treisman, A. (1969). Strategies and models of selective attention. *Psychological Review*, 76, 282 – 299.

Townsend, J. T. (1990). SERIAL VS PARALLEL PROCESSING: Sometimes They Look Like Tweedledum and Tweedledee but They Can (and Should) be Distinguished. *Psychological Science*, 1 (1), 16 – 54.

Tun, P. A., & Wingfield, A. (1994). Speech recall under heavy load conditions: Age, predictability, and limits on dual-task interference. *Aging, Neuropsychology, and Cognition*, 1(1), 29 – 44.

Turner, J. S. (1988). Design of a broadcast packet switching network. *IEEE Transactions on Communications*, 36 (6), 734 – 743.

Van Santen, J. P. H. (1993). Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language*, 7, 49 - 100.

Venkatagiri, H. S. (2003). Segmental intelligibility of four currently used text-to-speech synthesis methods. *The Journal of the Acoustical Society of America*, 113 (4), 2095 – 2104.

Villchur, E. (1973). Signal processing to improve speech intelligibility in perceptive deafness. *The Journal of the Acoustical Society of America*, 54 (1), 314 – 314.

Walden, B. E. (1984). Validity issues in speech recognition testing. In Glattke, T. J., & Elkins, E. (Eds.). *ASHA Reports Number 14. Speech recognition by the hearing impaired*. Rockville, MD: ASHA

Wang, W., Liew, S. C., & Li, V. O. K. (2005). Solutions to performance problems in VoIP over a 802.11 wireless LAN. *IEEE Transactions on Vehicular Technology*, 54 (1), 366 – 384.

Warren, R. M. (1970). Perceptual restorations of missing speech sounds. *Science*, 167, 392 – 393.

Whitaker, L. A., Peters, L. J., & Garinther, G. A. (1990). Effects of communication degradation on military crew task performance. In *Proceedings of symposium on psychology in the department of defense* (p. 274 – 278). Colorado Springs, CO: U.S. Air Force Academy.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance*. Hillsdale, NJ. Erlbaum

Wickens, C.D. (1984). Processing resources in attention. In Parasuraman, R., & Davies, D. R. (Eds.), *Varieties of attention*, New York, NY: Academic Press.

Wickens, C.D. (1992). *Engineering psychology and human performance* (2nd ed.). New York, NY: Harper Collins.

Wickens, C. D. (2000). *Imperfect and unreliable automation and its implication for attention allocation* (ARL-00-10/NASA-00-2). Urbana Champaign: University of Illinois.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance*. Englewood Cliffs, NJ: Prentice- Hall, Inc.

Wickens, C. D., Isreal, J. B., & Donchin, E. (1977). The event related cortical potential as an index of workload. *Proceedings of the Human Factors and Ergonomics Society*, 21, 282 – 287.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. In *Proceedings of the Human Factors Society 27th Annual Conference*, (pp.129 – 133). Santa Monica, CA: Human Factors Society

Winters, S. J., & Pisoni, D. B. (2004). Perception and comprehension of synthetic speech. In *Research on Spoken Language Processing Report*

No. 26 (pp. 97 – 138). Bloomington, IN: Speech Research Laboratory, Indiana University.

Yacullo, W. S., & Hawkins, D. B. (1987). Speech recognition in noise and reverberation by school age children. *Audiology*, 26, 235 – 246.

Yorkston, K. M., & Beukelman, D. R. (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders*, 46, 296 – 301.